2025-01-14

# Convergence analysis of real-time recurrent learning (RTRL) for a class of recurrent neural networks

# Convergence Analysis of Real-time Recurrent Learning (RTRL) for a class of Recurrent Neural Networks

Samuel Chun-Hei Lam*† , Justin Sirignano‡ , and Konstantinos Spiliopoulos§

January 15, 2025



**Abstract**

Recurrent neural networks (RNNs) are commonly trained with the truncated backpropagation-through-time (TBPTT) algorithm. For the purposes of computational tractability, the TBPTT algorithm truncates the chain rule and calculates the gradient on a finite block of the overall data sequence. Such approximation could lead to significant inaccuracies, as the block length for the truncated backpropagation is typically limited to be much smaller than the overall sequence length. In contrast, Real-time recurrent learning (RTRL) is an online optimization algorithm which asymptotically follows the true gradient of the loss on the data sequence as the number of sequence time steps $t \to \infty$. RTRL forward propagates the derivatives of the RNN hidden/memory units with respect to the parameters and, using the forward derivatives, performs online updates of the parameters at each time step in the data sequence. RTRL's online forward propagation allows for exact optimization over extremely long data sequences, although it can be computationally costly for models with large numbers of parameters. We prove convergence of the RTRL algorithm for a class of RNNs. The convergence analysis establishes a fixed point for the joint distribution of the data sequence, RNN hidden layer, and the RNN hidden layer forward derivatives as the number of data samples from the sequence and the number of training steps tend to infinity. We prove convergence of the RTRL algorithm to a stationary point of the loss. Numerical studies illustrate our theoretical results. One potential application area for RTRL is the analysis of financial data, which typically involve long time series and models with small to medium numbers of parameters. This makes RTRL computationally tractable and a potentially appealing optimization method for training models. Thus, we include an example of RTRL applied to limit order book data.

## 1 Introduction

Recurrent neural networks are commonly trained with the truncated *backpropagation* through time (TBPTT) algorithm, see for example [1, 2]. There is another training method which has recently received renewed interest: the *forward propagation* algorithm, frequently called "real-time recurrent learning" (RTRL) [3, 4, 5, 6, 7, 8]. In contrast to the TBPTT algorithm which uses backpropagation, the forward propagation algorithm solves an equation for the derivative of the hidden layer with respect to the parameters. The disadvantage of forward propagation is the computational cost will be $N \times d_\theta$ where $N$ is the number of hidden units and $d_\theta$ is the number of parameters. The advantage is that forward propagation can be used in an online algorithm to asymptotically optimize the RNN

*Author order is alphabetical.

†Mathematical Institute, University of Oxford Oxford, OX2 6GG, UK. Samuel.Lam@maths.ox.ac.uk

‡Mathematical Institute, University of Oxford Oxford, OX2 6GG, UK. Justin.Sirignano@maths.ox.ac.uk

§Department of Mathematics & Statistics, Boston University, Boston, MA, 02215. kspiliop@bu.edu

in the *true direction of steepest descent* for a long time series (i.e., as the number of time steps in the sequence $t \to \infty$) while TBPTT – due to its truncation of the chain rule after $\tau$ steps backward in time – is only an approximation (with potentially significant error).

Due to computational cost, $\tau$ will be limited to be much smaller than the overall length of a long data sequence. This is a significant limitation for TBPTT since it will lead to inaccurate estimates for the gradient of the loss. Therefore, for training RNNs (or other time series models) with small to moderate numbers of parameters on long data sequences, there is a strong argument that RTRL is superior to TBPTT. There has been recent interest in efficient numerical implementation and better understanding of RTRL; for example, see [9, 10, 11, 7, 12, 8]. Hybrid BPTT-RTRL schemes have also been proposed in the literature. The hybrid schemes are applicable when fully online learning is not the primary goal and can be used to compute untruncated gradients in an efficient way, see for example [13, 14, 8].

The RTRL algorithm forward propagates the derivatives of the RNN hidden/memory layer with respect to the parameters. Using these forward derivatives, which are updated at each time step of the sequence via a forward sensitivity equation, an online estimate for the gradient of the loss with respect to the parameters is calculated. This online gradient estimate is used at each time step to update the RNN parameters. As the number of data sequence time steps $t \to \infty$, the online estimate will converge to the true gradient of the loss for the data sequence. The RTRL algorithm is not limited by the length of the data sequence and can optimize over (extremely) long sequences.

In this paper, we study the convergence of the RTRL algorithm for a class of RNNs on long data sequences. We prove convergence of the RTRL algorithm as the number of data sequence time steps and parameter updates $t \to \infty$. The analysis requires addressing several mathematical challenges. First, the algorithm is fully online with simultaneous updates to the data samples, the RNN hidden layer, RNN hidden layer forward derivatives, and the parameters at each time step $t$. The data samples arrive from a data sequence and are correlated across time (they are not i.i.d. as in standard supervised learning frameworks). Analyzing the convergence of the training algorithm as $t \to \infty$ first requires establishing (joint) geometric ergodicity of the data sequence, RNN hidden layer, and RNN hidden layer forward derivatives. We use a fixed point analysis to prove a geometric convergence rate (uniform in the RNN parameters) to a unique stationary distribution. Using the geometric ergodicity and a Poisson equation, we can bound the fluctuations of the parameter evolution around the direction of steepest descent and prove convergence of the RNN parameters to a stationary point of the loss as $t \to \infty$.

We will now present the general RTRL algorithm which we study. Let $(X_t, Z_t, Y_t)$ be a geometrically ergodic process where $(X_t, Z_t)$ is a Markov chain, $Z_t$ is unobserved, $Y_t = f(X_t, Z_t, \eta_t)$, and $\eta_t$ is a random variable. A model is trained to predict $Y_t$ given the data sequence $(X_{t'})_{t' \le t}$. At each time step of the data sequence, the hidden layer $S_t^\theta \in \mathbb{R}^N$ of the RNN is updated:

$$S_{t+1}^\theta \quad = \quad g(S_t^\theta, X_t; \theta^S), \tag{1.1}$$

and the prediction for $Y_t$ is $f(S_t^\theta, X_t; \theta^f)$. The parameters that must be trained are $\theta = (\theta^S, \theta^f)$. The hidden layer $S_t^\theta$ is a nonlinear representation of the previous data sequence $(X_{t'})_{t' \le t}$. The update function $g(s, x; \theta^s)$ is determined by the parameters $\theta$ and therefore the hidden layer $S_t^\theta$ is a function of $\theta$.

Given $T$ observations in the sequence, the loss function is

$$L_T(\theta) = \frac{1}{T} \sum_{t=1}^{T} \ell\left(Y_t, f(S_t^\theta, X_t; \theta^f)\right), \tag{1.2}$$

where $\ell(y, f)$ is a loss function such as squared error or cross-entropy error. If $(S_t^\theta, X_t, Z_t, Y_t)$ is ergodic with stationary distribution $\pi_\theta(ds, dx, dz, dy)$, loss function $\mathbb{E}[L_T(\theta)] \to L(\theta)$ as $T \to \infty$

where

$$L(\theta) = \int \ell\big(y, f(s,x;\theta^f)\big)\,\pi_\theta(ds,dx,dz,dy). \tag{1.3}$$

In order to minimize $L(\theta)$, we evaluate its gradient

$$\begin{aligned}\nabla_\theta L(\theta) &= \nabla_\theta\bigg[\int \ell\big(y, f(s,x;\theta^f)\big)\pi_\theta(ds,dx,dz,dy)\bigg] \\ &= \nabla_\theta\bigg[\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\ell\big(Y_t, f(S_t^\theta, X_t;\theta^f)\big)\bigg].\end{aligned} \tag{1.4}$$

The stationary distribution $\pi_\theta(ds,dx,dz,dy)$ is unknown and therefore it is challenging to evaluate the gradient (1.4) in order to optimize the parameters $\theta$. The standard training method is TBPTT which, for the reasons previously discussed, does not use the true gradient.

Another approach would be to select a large $T$ and then use the standard gradient descent algorithm: (1): For a fixed $\theta$, simulate $(S_t^\theta, X_t, Z_t, Y_t)$ for $t = 0,1,\dots,T$, (2): Evaluate $\nabla_\theta L_T(\theta)$, (3): Update $\theta$ using gradient descent. However, this approach is prohibitively expensive. Since $T$ is large, a significant amount of computational time will be required in order to complete a single optimization iteration. Also, $\nabla_\theta L_T(\theta) \neq \nabla_\theta L(\theta)$, i.e. there will still be error even if $T$ is large.

Now, assuming we can interchange the gradient and limit (which will later be rigorously proven for the class of RNNs that we study),

$$\nabla_\theta L(\theta) = \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\nabla_\theta[\ell\big(Y_t, f(S_t^\theta, X_t;\theta^f)\big)]. \tag{1.5}$$

Define $\tilde{S}_t^\theta = \frac{\partial S_t^\theta}{\partial\theta}$. By chain rule, $\tilde{S}_t^\theta$ satisfies the forward propagation equation

$$\tilde{S}_{t+1}^\theta = \frac{\partial g}{\partial s}(S_t^\theta, X_t;\theta^S)\tilde{S}_t^\theta + \frac{\partial g}{\partial\theta}(S_t^\theta, X_t;\theta^S). \tag{1.6}$$

Therefore, we expect that we can write $\nabla_\theta L(\theta) = \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} G_t$ where

$$G_t = \frac{\partial\ell}{\partial f}\big(Y_t, f(S_t^\theta, X_t;\theta^f)\big)\bigg(\frac{\partial f}{\partial s}(S_t^\theta, X_t;\theta^f)\tilde{S}_t^\theta + \frac{\partial f}{\partial\theta}(S_t^\theta, X_t;\theta^f)\bigg). \tag{1.7}$$

This, therefore, motivates an *online forward propagation algorithm* for optimizing the RNN, commonly referred to as the RTRL algorithm, which trains the parameters $\theta_t = (\theta_t^S, \theta_t^f)$ according to:

$$\begin{aligned} S_{t+1} &= g(S_t, X_t;\theta_t^S), \\ \tilde{S}_{t+1} &= \frac{\partial g}{\partial s}(S_t, X_t;\theta_t^S)\tilde{S}_t + \frac{\partial g}{\partial\theta}(S_t, X_t;\theta_t^S), \\ G_t &= \frac{\partial\ell}{\partial f}\big(Y_t, f(S_t, X_t;\theta_t^f)\big)\bigg(\frac{\partial f}{\partial s}(S_t, X_t;\theta_t^f)\tilde{S}_t + \frac{\partial f}{\partial\theta}(S_t, X_t;\theta_t^f)\bigg), \\ \theta_{t+1} &= \theta_t - \alpha_t G_t, \end{aligned} \tag{1.8}$$

where the learning rate $\alpha_t$ is monotonically decreasing and satisfies the usual conditions $\sum_{t=0}^{\infty}\alpha_t = \infty$ and $\sum_{t=0}^{\infty}\alpha_t^2 < \infty$, see [15].

Note that the parameter $\theta$ is a fixed constant in the original equation for the RNN memory/hidden layer (1.1). The memory layer $S_t^\theta$ evolves for a constant $\theta$ which is fixed *a priori*. However, in the online forward propagation algorithm (1.8), $S_t, \tilde{S}_t$, and the parameter $\theta_t$ simultaneously evolve, which introduces complex dependencies and makes analysis of the algorithm challenging. The one-time-step transition probability function for $S_t, \tilde{S}_t$ will therefore change at every time step due to the evolution of the parameter $\theta_t$. For example, even if the function $g(\cdot)$ is linear (and then $S_t^\theta$ is a Gaussian process), $S_t$ in (1.8) will be non-Gaussian due to the joint dependence of the $(S_t, \theta_t)$ dynamics.

In this paper, we prove convergence of the RTRL algorithm (1.8) for a class of RNN models. To the best of our knowledge, our result is the first proof of *global* convergence for RTRL to a stationary point of its non-convex objective function. The only other existing mathematical analysis [16] that we are aware of studies *local* convergence of RTRL (i.e., if the parameters are initialized in a small neighborhood of the local minimizer) under *a priori* assumptions on the convergence rate of the RNN dynamics and the Hessian being positive definite at the local minimizer. We prove global convergence to a stationary point of the objective function (i.e., convergence for any parameter initialization) without assumptions on (A) the convergence rate of the RNN dynamics and (B) the Hessian structure. Our proof addresses (A) by using a fixed point analysis under the Wasserstein metric combined with a Poisson equation to quantify the speed of ergodicity. We address (B) using a cycle-of-stopping-times analysis to prove convergence to a stationary point of the objective function.

The rest of the paper is orgnanized as follows. The specific RNN architecture, assumptions, and main convergence theorem are presented in Section 2. In Section 3 we prove the geometric rate of convergence for the underlying sequences of interest. In Section 4 we study a Poisson equation that will be used to bound the fluctuation terms. We prove that the fluctuation term vanishes, at a suitably fast rate, as $t \to \infty$. The convergence as $t \to \infty$ of the RTRL algorithm is then established in Sections 5 and 6. In particular, in Section 5 we prove some necessary a-priori bounds which are then used in Section 6 to prove the convergence of the RTRL algorithm as $t \to \infty$. In Section 7 we present numerical studies illustrating our theoretical findings for several datasets. In particular, we compare the perfomrance of the TBPTT and RTRL algorithms. Appendix A contains detailed upper bound calculations for the derivatives of the sigmoidal activation function and Appendix B summarizes some important inequalities used throughout the paper.

# 2 RNN Model Architecture and Assumptions

## 2.1 Data generation

Let $d, N$ be *fixed* positive integers. Denote the input data as $X = (X_t)_{t\geq 0}$ with $X_t \in \mathbb{R}^d$ (as column vector), and the output data as $Y = (Y_t)_{t\geq 0}$ with $Y_t \in \mathbb{R}$ being a scalar. We assume that the input data could be lifted to a time-homogeneous Markov chain $(X_t, Z_t)_{t\geq 0}$ (for $Z_t \in \mathbb{R}^{d_Z}$) with transition kernel

$$\wp((x,z),A) = \mathbb{P}((X_{t+1}, Z_{t+1}) \in A \,|\, X_t = x, Z_t = z), \quad A \in \mathcal{B}(\mathbb{R}^{d+d_Z}). \tag{2.1}$$

The output data sequence $(Y_t)_{t\geq 0}$ depends on $(X_t, Z_t)$ as followed

$$Y_t = f(X_t, Z_t, \eta_t) \tag{2.2}$$

where $\eta_t \in \mathbb{R}$ are independent, identically distributed (iid) noises independent from the sequence $(X_t, Z_t)_{t\geq 0}$. We denote the probability distribution of $\eta_t$ as $\mu_\eta$. We note that the observation $(X_t, Y_t)_{t\geq 0}$ is not a Markov process in general.

Next, we define the $p$-Wasserstein distance, which will be used later, in particularly for $p = 2$:

**Definition 2.1** ($p$-Wasserstein distance)**.** Let $(\mathcal{X}, \|\cdot\|)$ be a Polish (complete separable) normed space, and $\mathcal{M}(\mathcal{X})$ be the space of all measures on $\mathcal{X}$. We define the space of measures on $\mathcal{X}$ with $p$-th moment as

$$\mathcal{P}_p(\mathcal{X}) = \left\{ \mu \in \mathcal{M}(\mathcal{X}) \;\middle|\; \int_{\mathcal{X}} \|\mathsf{x}\|^p \, \mu(d\mathsf{x}) < +\infty \right\}. \tag{2.3}$$

The $p$-Wasserstein distance between $\mu, \nu \in \mathcal{P}_p(\mathcal{X})$ is

$$\mathsf{Wass}_p(\mu, \nu) = \inf_{\gamma} \left( \int_{\mathcal{X}\times\mathcal{X}} \|\mathsf{x} - \tilde{\mathsf{x}}\|^p \, \gamma(d\mathsf{x}, d\tilde{\mathsf{x}}) \right)^{1/p}, \tag{2.4}$$

where $\gamma$ is a coupling of $\mu, \nu$, such that $\gamma(A \times \mathcal{X}) = \mu(A)$ and $\gamma(\mathcal{X} \times B) = \nu(B)$.

*Remark* 2.2. We note that $\mathcal{P}_p(\mathcal{X})$ is Polish with respect to $\mathsf{Wass}_p$. Moreover, by e.g. [17, Theorem 4.1], there exists an *optimal* coupling $\gamma^*$ such that

$$[\mathsf{Wass}_p(\mu, \nu)]^p = \int_{\mathcal{X}\times\mathcal{X}} \|\mathsf{x} - \tilde{\mathsf{x}}\|^p \, \gamma^*(d\mathsf{x}, d\tilde{\mathsf{x}}). \tag{2.5}$$

Here, we will let $\mathcal{X}$ be a finite-dimensional Euclidean space, and $\|\cdot\|$ being the max-norm $\|x\|_{\max} = \max_i |x_i|$. Analogously, in the case of a matrix $M = [M_{i,j}]_{i,j}$ in finite dimensions we define $\|M\|_{\max} = \max_{i,j} |M_{i,j}|$. Finally, $\mathcal{P}_p(\mathcal{X})$ equipped with the Wasserstein metric $\mathsf{Wass}_p(\cdot,\cdot)$ is Polish as $\mathcal{X}$ is Polish (see e.g. [18, Theorem 1.1.3]).

With the notion of $p$-Wasserstein distance, we state the assumptions on the sequences $(X_t, Z_t)_{t\geq 0}$ and the output sequence $(Y_t)_{t\geq 0}$.

**Assumption 2.3** (On the dynamics and background noises of the data sequences)**.**

1. We assume that the transitional kernel $\wp$ is $L_\wp$-contractive with respect to max-Wasserstein distance for $L_\wp < 1$. This means that

$$\sup_{(x,z)\neq(\tilde{x},\tilde{z})} \frac{\mathsf{Wass}_2(\wp((x,z),\cdot), \wp((\tilde{x},\tilde{z}),\cdot))}{\|(x,z) - (\tilde{x},\tilde{z})\|_{\max}} \leq L_\wp < 1. \tag{2.6}$$

2. We assume that the function $f$ is $L_f$-globally Lipschitz, that is for any $(x,z,\eta), (\tilde{x},\tilde{z},\tilde{\eta}) \in \mathbb{R}^{d+d_Z+1}$ and $\eta$ we have

$$|f(x,z,\eta) - f(\tilde{x},\tilde{z},\eta)| \leq L_f \, \|(x,z,\eta) - (\tilde{x},\tilde{z},\tilde{\eta})\|_{\max}. \tag{2.7}$$

3. We finally assume that the sequence $(X_t, Z_t)_{t\geq 0}$ is bounded with $\|(X_t, Z_t)\|_{\max} \leq 1$, and that $f$ is bounded by the constant $L_f > 0$.

An example of a Markov chain $(X_t, Z_t)_{t\geq 0}$ that satisfies the above assumption is the following:

**Example 2.4.** Let $g : \mathbb{R}^{d+d_Z} \to \mathbb{R}^{d+d_Z}$ be a $L_\wp$-Lipschitz function bounded by $1/2$, and that $(\epsilon_k)_{k\geq 0}$ is an iid mean zero noise bounded by $1/2$ with distribution $\mu_\epsilon$. The chain $(X_t, Z_t)$ defined recursively by

$$(X_{t+1}, Z_{t+1}) = g(X_k, Z_k) + \epsilon_k, \quad \|(X_0, Z_0)\|_{\max} \leq 1, \tag{2.8}$$

automatically satisfy the boundedness assumption $\|(X_t, Z_t)\|_{\max} \le 1$. Moreover, one could construct a measure $\gamma_{(\mathsf{x},\mathsf{z}),(\tilde{\mathsf{x}},\tilde{\mathsf{z}})}(\cdot)$ on $\mathbb{R}^{d+1} \times \mathbb{R}^{d+1}$, such that

$$\gamma_{(x,z),(\tilde{x},\tilde{z})}(A \times B) = \int_{\mathbb{R}^{d+1}\times\mathbb{R}^{d+1}} \mathbb{I}_A(g(x,z)+\epsilon)\mathbb{I}_B(g(\tilde{x},\tilde{z})+\epsilon)\,\mu_\epsilon(d\epsilon). \tag{2.9}$$

With this,

$$\begin{aligned}(\mathsf{Wass}_2(\wp((x,z),\cdot),\wp((\tilde{x},\tilde{z}),\cdot)))^2 &\le \int_{\mathbb{R}^{d+1}\times\mathbb{R}^{d+1}} \|(\mathsf{x},\mathsf{z})-(\tilde{\mathsf{x}},\tilde{\mathsf{z}})\|^2_{\max}\,\gamma_{(x,z),(\tilde{x},\tilde{z})}(d\mathsf{x},d\mathsf{z},d\tilde{\mathsf{x}},d\tilde{\mathsf{z}}) \\ &\le \int_{\mathbb{R}^{d+1}\times\mathbb{R}^{d+1}} \|g(x,z)+\epsilon-g(\tilde{x},\tilde{z})-\epsilon\|^2_{\max}\,\gamma_{(x,z),(\tilde{x},\tilde{z})}(d\mathsf{x},d\mathsf{z},d\tilde{\mathsf{x}},d\tilde{\mathsf{z}}) \\ &\le L_\wp^2\|(x,z)-(\tilde{x},\tilde{z})\|^2_{\max}. \end{aligned} \tag{2.10}$$

Hence

$$\sup_{(x,z)\neq(\tilde{x},\tilde{z})} \frac{\mathsf{Wass}_2(\wp((x,z),\cdot),\wp((\tilde{x},\tilde{z}),\cdot))}{\|(x,z)-(\tilde{x},\tilde{z})\|_{\max}} \le L_\wp < 1. \tag{2.11}$$

*Remark* 2.5 (Ergodicity of the data sequence). We note that the Markov chain $(X_k, Z_k)_{k\ge1}$ is *geometrically ergodic*. In particular, if we let $\rho_k$ be the distribution of $(X_k, Z_k)$ at time step $k$, then we can write $\rho_k := (\wp^\vee)^k\rho_0 = (\wp^k)^\vee\rho_0$, where for probability measures $\rho_0$ and $A \in \mathcal{B}(\mathbb{R}^{d+d_Z})$,

$$\wp^\vee\rho(A) := \int_{\mathbb{R}^{d+d_Z}} \wp(x,A)\,\rho(dx), \tag{2.12}$$

and the $k$-step transition probabilities $\wp^k$ defined iteratively as

$$\wp^{k+1}((x,z),A) = \int_A \wp((\mathsf{x},\mathsf{z}),A)\,\wp^k((x,z),d\mathsf{x}\,d\mathsf{z}), \quad \wp^0((x,z),A) = \delta_{(x,z)}(A). \tag{2.13}$$

By [19, Proposition 14.3], for any probability measures $\rho, \tilde{\rho}$

$$\mathsf{Wass}_2(\wp^\vee\rho,\wp^\vee\tilde{\rho}) \le L_\wp \mathsf{Wass}_2(\rho,\tilde{\rho}), \tag{2.14}$$

and so by Banach Fixed Point Theorem, there exists a unique probability measure $\rho^*$ such that

$$\wp^\vee\rho^* = \rho^*, \tag{2.15}$$

(i.e., $\rho^*$ is a stationary measure of the Markov chain $(X_t, Z_t)$), and that

$$\mathsf{Wass}_2((\wp^\vee)^t\rho,\rho^*) \le \frac{L_\wp^t}{1-L_\wp}\mathsf{Wass}_2(\wp^\vee\rho,\,\rho). \tag{2.16}$$

## 2.2 Recurrent Neural Network

In this subsection we fix notation in handling recurrent neural networks, paying attention to the convention of flattening tensors to matrices, and matrices to vectors.

The recurrent neural network is governed by the parameters $\theta = (A, W, B, c)$ with $A \in \mathbb{R}^{N\times d}$, $W \in \mathbb{R}^{N\times N}$, $B \in \mathbb{R}^N$ and $c \in \mathbb{R}$. We shall adopt the following slicing notations:

**Notation 2.6** (Slicing notation)**.** We adopt the standard slicing notation, so that given vector $B \in \mathbb{R}^d$ and matrix $A \in \mathbb{R}^{N\times d}$,

- $B^i$ or $[B]_i$ refers to the $i$-th entry of the vector $B$,
- $A^{i,:}$, view as a $\mathbb{R}^d$ vector, refers to the $i$-th row of the matrix $A$,
- $A^{:,j}$, view as a $\mathbb{R}^N$ vector, refers to the $j$-th column of the matrix $A$, and
- $A^{i,j}$ or $A^{ij}$ or $[A]_{ij}$ refers to the $(i,j)$-th entry of the matrix $A$.

Given the input data $X = (X_t)_{t\geq 0}$, the output of the recurrent neural network (RNN) is modeled as

$$\begin{aligned} S^i_{t+1} = S^i_{t+1}(X;\theta) &:= \sigma\left(\frac{1}{d}\sum_{j=1}^{d}\phi(A^{ij})X^j_t + \frac{1}{N}\sum_{j=1}^{N}\phi(W^{ij})S^j_t\right), \\ &= \sigma\left(\frac{1}{d}\left\langle\phi(A^{i,:}), X_t\right\rangle + \frac{1}{N}\left\langle\phi(W^{i,:}), S_t\right\rangle\right) \end{aligned} \tag{2.17}$$

$$\hat{Y}_t = f_t(X;\theta) := \langle\lambda(B), S_t\rangle + \lambda(c) = \sum_{i=1}^{N}\lambda(B^i)S^i_t + \lambda(c), \tag{2.18}$$

where $\sigma$ is the standard sigmoid function

$$\sigma(z) = \frac{1}{1+e^{-z}}; \tag{2.19}$$

and $\lambda, \phi \in C^\infty_b$ are two infinitely differentiable function with $|\lambda| \leq C_\lambda$ and $|\phi| \leq C_\phi$ for appropriate $C_\lambda, C_\phi > 0$. We adopt the usual convention of $\lambda(B) = (\lambda(B^1), ..., \lambda(B^N))$ for $B \in \mathbb{R}^N$ and similarly for $\lambda(C)$, $\phi(A)$ and $\phi(B)$.

*Remark* 2.7. We note here that the constant $C_\phi$ is selected, via Assumption 2.10, to be small enough to ensure a fixed point in our analysis. On the other hand, for $\lambda$ we only need to know that $C_\lambda < \infty$. Introducing the functions $\phi, \lambda$ can be viewed as a smooth way to constraint the phase space of the parameters, which is important for the fixed point analysis of this paper to go through.

*Remark* 2.8. We emphasize that the number of hidden units, $N$, is not to be send to $\infty$ in our analysis.

The sigmoid function is used as its derivatives are uniformly bounded. In particular,

**Lemma 2.9.** *[Bounds for the derivatives of the sigmoid function] Let $\sigma : \mathbb{R} \to \mathbb{R}$ be the usual sigmoid function $\sigma(y) = (1+e^{-y})^{-1}$. Then*

$$\max_y |\sigma(y)| < 1, \quad \max_y |\sigma'(y)| \leq \frac{1}{4}, \quad \max_y |\sigma''(y)| \leq \frac{\sqrt{3}}{18} < \frac{1}{10}, \quad \max_y |\sigma'''(y)| \leq \frac{1}{8}.$$

*Proof.* See Section A. □

The following assumptions are also necessary for the fixed point analysis:

**Assumption 2.10.** Assume that the functions $\phi, \lambda$ are bounded, at least twice continuously differentiable with all derivatives bounded. In addition, letting $M_\phi = 1 + [C_\phi/(4 - C_\phi)]$,

$$
\begin{aligned}
L_0 &= \max\left(\frac{3C_{\phi'}}{\min(N,d)}\left(\frac{C_\phi M_\phi}{10} + \frac{1}{4}\right), \frac{3C_\phi}{4}\right) \\
L_1 &= 4\max\left[\frac{M_\phi C_\phi^2}{5(\min(d,N))^2} + \frac{C_\phi M_\phi^2 C_{\phi'}^2}{8(\min(d,N))^2} + \frac{C_{\phi''}}{4\min(N,d)}\right. \\
&\quad \left. + \frac{C_\phi}{10}\left(L_2 + \frac{C_\phi}{1 - C_\phi/4}\left(\frac{C_\phi^2 M_\phi^2}{10(\min(N,d))^2} + \frac{1}{4}L_2\right)\right), \frac{C_\phi M_\phi C_{\phi'}}{5\min(N,d)} + \frac{C_{\phi'}}{4N}\right] \vee C_\phi, \\
L_2 &= \frac{C_{\phi''}}{\min(N,d)} \vee \frac{C_{\phi'}^2}{Nd(4 - C_\phi)}.
\end{aligned}
$$

Let further $L = L_0 \vee L_1$ We shall assume that $0 < q := \sqrt{L^2 + L_\wp^2} < 1$.

Note that Assumption 2.10 is achievable for example by functions $\phi$ that are sufficiently bounded. For instance, one could consider suitably scaled sigmoid functions.

**Notation 2.11** (Vectorising the parameters)**.** Note that the parameter $\theta$ could be viewed as a vector in $\Theta = \mathbb{R}^{d_\theta}$, where $d_\theta = Nd + N^2 + N + d$. We shall arrange the entries of the parameter as followed:

$$
\theta = \begin{bmatrix} \mathsf{vec}(A) \\ \mathsf{vec}(W) \\ B \\ c \end{bmatrix},
$$

where $\mathsf{vec}(A)$ is the vectorisation of the matrix $A$, formulated by stacking the *columns* of $A$:

$$
\mathsf{vec}(A) = \begin{bmatrix} A_{:,1} \\ \vdots \\ A_{:,d} \end{bmatrix} = \begin{bmatrix} A_{11} \\ A_{21} \\ \vdots \\ A_{n1} \\ \vdots \\ A_{1d} \\ \vdots \\ A_{nd} \end{bmatrix}.
$$

We shall use $A$ to denote both the matrix itself and its vectorisation when the context is clear.

### 2.2.1 Gradient of the loss function

The parameters are selected to minimize the loss function

$$
L_T(\theta) = \frac{1}{2T}\sum_{t=1}^{T}\left(Y_t - f_t(\theta)\right)^2.
$$

To compute the derivatives of the loss function, it is important to study the forward derivatives of the memory process $S_t$:

$$
\tilde{S}_t^{\diamondsuit} = \frac{\partial S_t}{\partial \diamondsuit},
$$

where $\diamond$ represents one or more parameter(s) of the neural network, e.g. an entry of the weight matrix $A^{ij}$, the full weight matrix $A$, and the vector of all parameters $\theta$. We are particularly interested with the following forward derivatives:

$$\tilde{S}_t^A = \frac{\partial S_t}{\partial A}, \quad \tilde{S}_t^W = \frac{\partial S_t}{\partial W}.$$

They shall be indexed in the following way:

$$\tilde{S}_t^{A^{ij},:} = \frac{\partial S_t}{\partial A^{ij}}, \quad \tilde{S}_t^{A^{ij},k} = \frac{\partial S_t^k}{\partial A^{ij}}, \quad \tilde{S}_t^{W^{ij},:} = \frac{\partial S_t}{\partial W^{ij}}, \quad \tilde{S}_t^{W^{ij},k} = \frac{\partial S_t^k}{\partial W^{ij}}.$$

Using chain rule, we can compute an evolution equation of the forward derivatives:

$$\tilde{S}_{t+1}^{A^{ij},k} = \sigma'\left(\frac{1}{d}\left\langle \phi(A^{k,:}), X_t\right\rangle + \frac{1}{N}\left\langle \phi(W^{k,:}), S_t\right\rangle\right) \times \left(\frac{1}{d}\delta_{ik}\phi'(A^{ij})X_t^j + \frac{1}{N}\sum_{\ell=1}^{N}\phi(W^{k\ell})\tilde{S}_t^{A^{ij},\ell}\right), \quad (2.20)$$

$$\tilde{S}_{t+1}^{W^{ij},k} = \sigma'\left(\frac{1}{d}\left\langle \phi(A^{k,:}), X_t\right\rangle + \frac{1}{N}\left\langle \phi(W^{k,:}), S_t\right\rangle\right) \times \left(\frac{1}{N}\delta_{ik}\phi'(W^{ij})S_t^j + \frac{1}{N}\sum_{\ell=1}^{N}\phi(W^{k\ell})\tilde{S}_t^{W^{ij},\ell}\right), \quad (2.21)$$

where $\delta_{ik} = 1$ if $i = k$ and zero otherwise.

**Notation 2.12** (Flattening the first derivatives)**.** Depending on what $\diamond$ is, $\tilde{S}_t^{\diamond}$ could be a scalar, vector, matrix or rank-3 tensor. For example, $\tilde{S}_t^A$ and $\tilde{S}_t^W$ are both rank-3 tensors.

It is sometimes convenient to flatten $\tilde{S}_t^A$ into a $(N \times Nd)$-dimensional matrix with entries arranged as follows:

$$\mathsf{flat}(\tilde{S}_t^A) = \begin{bmatrix} \uparrow & & \uparrow & & \uparrow & & \uparrow \\ \frac{\partial S_t}{\partial A^{11}} & \cdots & \frac{\partial S_t}{\partial A^{n1}} & \cdots & \frac{\partial S_t}{\partial A^{1d}} & \cdots & \frac{\partial S_t}{\partial A^{nd}} \\ \downarrow & & \downarrow & & \downarrow & & \downarrow \end{bmatrix}$$
$$= \begin{bmatrix} \tilde{S}_t^{A^{11},1} & \cdots & \tilde{S}_t^{A^{n1},1} & \cdots & \tilde{S}_t^{A^{1d},1} & \cdots & \tilde{S}_t^{A^{nd},1} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \tilde{S}_t^{A^{11},k} & \cdots & \tilde{S}_t^{A^{n1},k} & \cdots & \tilde{S}_t^{A^{1d},k} & \cdots & \tilde{S}_t^{A^{nd},k} \end{bmatrix},$$

and similarly $\tilde{S}_t^W$ into a $(N \times N^2)$-matrix.

We shall also refer $\tilde{S}^{A,k}$ as the first derivative of memory neuron $S^k$ with respect to all parameters in matrix $A$, which is itself a matrix. Similarly for $\tilde{S}^{W,k}$.

In this paper, we shall abuse notation and not indicate the vectorising/flattening operations of the matrices/vectors when the context is clear.

We shall now compute the derivatives of the loss function. We start by the following application of the chain rule:

$$\frac{\partial L_T}{\partial \diamond} = -\frac{1}{T}\sum_{t=1}^{T}(Y_t - f_t(\theta))\frac{\partial f_t}{\partial \diamond}. \quad (2.22)$$

It is therefore crucial to compute the gradients $\partial f_t/\partial\diamond$.

**Notation 2.13** (Vectorising the gradients)**.** For any functions $g$ that depends on $\theta$, $\nabla_\diamond g$ will *always* refer to the *vectorised* version of $\partial g/\partial \diamond$, or simply $\partial g/\partial \mathsf{vec}(\diamond)$. For example, we have

$$\nabla_A g = \begin{bmatrix} \partial g/\partial A^{11} \\ \vdots \\ \partial g/\partial A^{N1} \\ \vdots \\ \partial g/\partial A^{1d} \\ \vdots \\ \partial g/\partial A^{Nd} \end{bmatrix},$$

and similarly for $\nabla_W g$. Finally, we denote $\nabla_\theta g$ as the (vectorised) gradient of $g$ with respect to all parameters:

$$\nabla_\theta g = \begin{bmatrix} \nabla_A g \\ \nabla_W g \\ \nabla_B g \\ \partial g/\partial c \end{bmatrix}.$$

It is easy to show that

$$\nabla_B f_t = \lambda'(B) \odot S_t, \quad \partial f_t/\partial c = \lambda'(c),$$

where $\odot$ denotes the element-wise multiplication (Hadarmard product), such that if $u, v \in \mathbb{R}^d$ then $u \odot v$ is also a vector in $\mathbb{R}^d$ with entries $(u \odot v)^i = u^i v^i$.

The gradient of the loss function is given by

$$\nabla_\theta L_T(\theta) = \frac{1}{T}\sum_{t=1}^{T} \tilde{G}_t, \quad \tilde{G}_t = -(Y_t - f_t(\theta)) \times \begin{bmatrix} \sum_k \lambda(B^k)\, \mathsf{vec}\left(\tilde{S}_t^{A,k}\right) \\ \sum_k \lambda(B^k)\, \mathsf{vec}\left(\tilde{S}_t^{W,k}\right) \\ \lambda'(B) \odot S_t \\ \lambda'(c) \end{bmatrix}. \tag{2.23}$$

We observe that $\tilde{G}_t$ is a stochatic biased estimate of $\nabla_\theta L_T(\theta)$. This is crucial in developing the online forward SGD used for minimising the loss function.

## 2.3 Online Forward SGD

We will now introduce the online forward stochastic gradient descent algorithm to train the parameters $\theta$. At each time step $t$, the parameters will be updated. The current iteration of the trained parameters at time $t$ will be denoted $\theta_t = \{A_t, W_t, B_t, c_t\}$. An online estimate of the forward gradient will also be introduced:

$$\begin{aligned} \hat{S}_{t+1}^{A^{ij},k} &= \sigma'\left(\frac{1}{d}\left\langle \phi(A_t^{k,:}), X_t \right\rangle + \frac{1}{N}\left\langle \phi(W_t^{k,:}), \bar{S}_t \right\rangle\right) \times \left(\frac{1}{d}\delta_{ik}\phi'(A_t^{ij})X_t^j + \frac{1}{N}\sum_{\ell=1}^{N}\phi(W_t^{k\ell})\hat{S}_t^{A_{ij},\ell}\right), \\ \hat{S}_{t+1}^{W^{ij},k} &= \sigma'\left(\frac{1}{d}\left\langle \phi(A_t^{k,:}), X_t \right\rangle + \frac{1}{N}\left\langle \phi(W_t^{k,:}), \bar{S}_t \right\rangle\right) \times \left(\frac{1}{N}\delta_{ik}\phi'(W_t^{ij})\bar{S}_t^j + \frac{1}{N}\sum_{\ell=1}^{N}\phi(W_t^{k\ell})\hat{S}_t^{W_{ij},\ell}\right), \end{aligned} \tag{2.24}$$

where

$$\begin{aligned} \bar{S}_{t+1}^i &= \sigma\left(\frac{1}{d}\left\langle \phi(A_t^{i,:}), X_t \right\rangle + \frac{1}{N}\left\langle \phi(W_t^{i,:}), \bar{S}_t \right\rangle\right) = \sigma\left(\frac{1}{d}\sum_{j=1}^{d}\phi(A^{ij})X_t^j + \frac{1}{N}\sum_{j=1}^{N}\phi(W_t^{ij})\bar{S}_t^j\right), \\ \bar{f}_t(\theta) &= \left\langle \lambda(B_t), \bar{S}_t \right\rangle + \lambda(c_t) = \sum_{j=1}^{N}\lambda(B_t^j)\bar{S}_t^j + \lambda(c_t). \end{aligned}$$

Note that $\bar{S}_t$ is different than $S_t$ since its updates are governed by the evolving, trained parameters $\theta_t$ instead of a fixed constant parameter $\theta$.

We will approximate the gradients with these online stochastic estimates:

$$\nabla_\theta L_T(\theta_t) \approx G_t := -(Y_t - \bar{f}_t(\theta_t)) \times \begin{bmatrix} \sum_k \lambda(B_t^k)\mathsf{vec}(\hat{S}_t^{A,k}) \\ \sum_k \lambda(B_t^k)\mathsf{vec}(\hat{S}_t^{W,k}) \\ \lambda'(B_t) \odot \bar{S}_t \\ \lambda'(c_t) \end{bmatrix}. \tag{2.25}$$

The *forward* stochastic gradient descent algorithm is:

$$\theta_{t+1} = \theta_t - \alpha_t G_t, \tag{2.26}$$

where the learning rate $\alpha_t$ satisfies the standard conditions:

**Assumption 2.14.** We assume that the learning rates are monotonic decreasing and satisfy the Robin-Munro conditions

$$\sum_{t\geq 0} \alpha_t = \infty, \quad \sum_{t\geq 0} \alpha_t^2 < \infty.$$

*Remark* 2.15. We make a short remark that if $\sum_{t\geq 0} \alpha_t^2 \leq M < +\infty$, then

$$\sum_{t\geq 0} \alpha_{t+1}\alpha_t \leq \left(\sum_{t\geq 0} \alpha_t^2\right)^{1/2} \left(\sum_{t\geq 0} \alpha_{t+1}^2\right)^{1/2} \leq \sum_{t\geq 0} \alpha_t^2 \leq M < +\infty.$$

Let $\mu_t^\theta$ be the distribution of $(S_t, \tilde{S}_t^\theta, Y_t, X_t, Z_t)$. In Section 3, we will prove that $\mu_t^\theta$ converges to a unique stationary distribution $\mu^\theta$ as $t \to \infty$ in the Wasserstein metric $W_1$. For long time series as $t \to \infty$, the loss function will therefore become

$$L_T(\theta) \overset{L^2}{\to} L(\theta) = \mathbb{E}\left[\frac{1}{2}\left(Y - f(\theta)\right)^2\right], \tag{2.27}$$

where $f(\theta) = \lambda(B)S + \lambda(C)X$ and $(S, \tilde{S}, Y, X, Z) \sim \mu^\theta$. This result also relies upon the fact that $\tilde{S}_t^\theta$ is bounded, which we will prove later, and $S_t, Y_t, X_t, Z_t$, by definition, are bounded.

The parameter evolution can be decomposed into the direction of steepest descent and a fluctuation term:

$$\theta_{t+1} = \theta_t \underbrace{-\alpha_t \nabla_\theta L(\theta_t)}_{\text{main descent term}} \underbrace{-\alpha_t(G_t - \nabla_\theta L(\theta_t))}_{\text{fluctuations term}}. \tag{2.28}$$

In what follows we establish in Theorem 6.1 that, under Assumptions 2.3 and 2.10 we have

$$\lim_{t\to\infty} \mathbb{E}\|\nabla_\theta L(\theta_t)\|_2 = 0,$$

where $\|\cdot\|_2$ denotes the standard $\ell_2$–Euclidean norm of a vector. The proof of this convergence result for the RTRL algorithm is based on a detailed analysis of how $\theta_t$ and the fluctuations term $\alpha_t(G_t - \nabla_\theta L(\theta_t))$ behave as $t \to \infty$.

### 2.4 Forward Second Derivatives

In proving the stability the RTRL we shall look into the second derivatives (Hessian) of the recurrent neural network. The Hessian is a tensor, so we shall fix the convention for flattening it.

We shall let $\tilde{\tilde{S}}_t$ be the Hessian of $S_t$ with respect to the weight matrices $A$ and $W$. This is a rank-5 tensor, and we shall index it by the following:

$$\tilde{\tilde{S}}_t^{\diamond,\heartsuit,k} = \frac{\partial S_t^k}{\partial \diamond \partial \heartsuit}, \quad \tilde{\tilde{S}}_t^{\diamond,\heartsuit,:} = \frac{\partial S_t}{\partial \diamond \partial \heartsuit}$$

where $\diamond$, $\heartsuit$ are any entries in either matrix $A$ or matrix $W$. For example, we shall have

$$\tilde{\tilde{S}}_t^{A^{mn},A^{ij},:} = \frac{\partial S_t}{\partial A^{mn} \partial A^{ij}}, \quad \tilde{\tilde{S}}_{t+1}^{A^{mn},W^{ij},:} = \frac{\partial S_t}{\partial A^{mn} \partial W^{ij}}, \quad \tilde{\tilde{S}}_{t+1}^{W^{mn},W^{ij},:} = \frac{\partial S_t}{\partial W^{mn} \partial W^{ij}},$$
$$\tilde{\tilde{S}}_t^{A^{mn},A^{ij},k} = \frac{\partial S_t^k}{\partial A^{mn} \partial A^{ij}}, \quad \tilde{\tilde{S}}_{t+1}^{A^{mn},W^{ij},k} = \frac{\partial S_t^k}{\partial A^{mn} \partial W^{ij}}, \quad \tilde{\tilde{S}}_{t+1}^{W^{mn},W^{ij},k} = \frac{\partial S_t^k}{\partial W^{mn} \partial W^{ij}}.$$

We shall split the $\tilde{\tilde{S}}_t$ into the following components:

- $\tilde{\tilde{S}}_t^{AA}$, that consists entries of the form $\tilde{\tilde{S}}_t^{A^{mn},A^{ij},:}$,
- $\tilde{\tilde{S}}_t^{AW}$, that consists entries of the form $\tilde{\tilde{S}}_t^{A^{mn},W^{ij},:}$,
- $\tilde{\tilde{S}}_t^{WA}$, that consists entries of the form $\tilde{\tilde{S}}_t^{W^{mn},A^{ij},:}$, and
- $\tilde{\tilde{S}}_t^{WW}$, that consists entries of the form $\tilde{\tilde{S}}_t^{W^{mn},W^{ij},:}$.

All these components are five-dimensional tensor.

We shall flatten each of the components to three-dimensional tensors, such that the *frontal* slides are defined as follow: for $\tilde{\tilde{S}}_t^{AA}$ we have

$$[\mathsf{flat}(\tilde{\tilde{S}}_t^{AA,k})] = \begin{bmatrix} \tilde{\tilde{S}}^{A_{11},A_{11},k} & \dots & \tilde{\tilde{S}}^{A_{11},A_{n1},k} & \dots & \tilde{\tilde{S}}^{A_{11},A_{1d},k} & \dots & \tilde{\tilde{S}}^{A_{11},A_{nd},k} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \tilde{\tilde{S}}^{A_{n1},A_{11},k} & \dots & \tilde{\tilde{S}}^{A_{n1},A_{n1},k} & \dots & \tilde{\tilde{S}}^{A_{n1},A_{1d},k} & \dots & \tilde{\tilde{S}}^{A_{n1},A_{nd},k} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \tilde{\tilde{S}}^{A_{1d},A_{11},k} & \dots & \tilde{\tilde{S}}^{A_{1d},A_{n1},k} & \dots & \tilde{\tilde{S}}^{A_{1d},A_{1d},k} & \dots & \tilde{\tilde{S}}^{A_{1d},A_{nd},k} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \tilde{\tilde{S}}^{A_{nd},A_{11},k} & \dots & \tilde{\tilde{S}}^{A_{nd},A_{n1},k} & \dots & \tilde{\tilde{S}}^{A_{nd},A_{1d},k} & \dots & \tilde{\tilde{S}}^{A_{nd},A_{nd},k} \end{bmatrix},$$

and similar for the other components. We note that the following symmetry holds for the frontal slides of the other flatten tensors:

$$\mathsf{flat}(\tilde{\tilde{S}}_t^{AA,k})^\top = \mathsf{flat}(\tilde{\tilde{S}}_t^{AA,k}), \quad \mathsf{flat}(\tilde{\tilde{S}}_t^{WW,k})^\top = \mathsf{flat}(\tilde{\tilde{S}}_t^{WW,k}), \quad \mathsf{flat}(\tilde{\tilde{S}}_t^{AW,k})^\top_{:,:,k} = \mathsf{flat}(\tilde{\tilde{S}}_t^{WA}).$$

The $\tilde{\tilde{S}}_t$ shall also be flattened such that the following holds:

$$[\mathsf{flat}(\tilde{\tilde{S}}_t)]_{:,:,k} = \begin{bmatrix} [\mathsf{flat}(\tilde{\tilde{S}}_t^{AA})]_{:,:,k} & [\mathsf{flat}(\tilde{\tilde{S}}_t^{AW})]_{:,:,k} \\ [\mathsf{flat}(\tilde{\tilde{S}}_t^{WA})]_{:,:,k} & [\mathsf{flat}(\tilde{\tilde{S}}_t^{WW})]_{:,:,k} \end{bmatrix}.$$

As for the case of first derivatives, we shall also use $\tilde{\tilde{S}}^{AA,k}$ as the 3D tensor $[\tilde{\tilde{S}}_t^{AA}]_{:,:,k}$, and similarly for other second derivatives as well. Finally, we may vectorise $\tilde{\tilde{S}}_t$ from the flattened version of $\tilde{\tilde{S}}_t$.

# 3 Fixed Point Analysis

We will study the convergence in distribution of the joint process of input sequence, value of hidden layers and the derivatives of hidden layers with respect to the parameters. For convenience we shall also take the noise $\eta_t$ into consideration. The joint process is, hence, in the form

$$(X_t, Z_t, S_t, \tilde{S}_t, \tilde{\tilde{S}}_t, \eta_t), \tag{3.1}$$

where

- $(X_t, Z_t)$ is the input sequence as defined in assumption 2.3,
- $S_t$ is the hidden layers of the Recurrent Neural Network, as defined in the previous section,
- $\tilde{S}_t = (\nabla_A S_t, \nabla_W S_t)$ is the first derivatives (or gradient) of the hidden layers with respect to the parameters $A$ and $W$, as computed in equation (2.21)
- $\tilde{\tilde{S}}_t = \nabla^2_{A,W} S_t$ is the second derivatives (or Hessian) of the hidden layers with respect to the parameters $A$ and $W$.

This is a process on the state space

$$\mathcal{X} = \mathbb{R}^{N_H}, \quad N_H = d + d_Z + N^2(d+N) + N^3(d^2 + Nd + N^2) + 1.$$

The main objective is to show that $H_t$ admits a random fixed point.

## 3.1 An intermediate lemma for fixed point analysis

Given a general Markov chain $(U_t)_{t\geq 0}$ in $\mathbb{R}^{N_1}$ with transition kernel $\wp$ that satisfies the contraction assumption

$$\frac{\mathsf{Wass}_2(\wp(u,\cdot), \wp(\tilde{u},\cdot))}{\|u-\tilde{u}\|_{\max}} \leq L_\wp < 1, \tag{3.2}$$

where the Wasserstein distance is defined with respect to the max norm. Define the random sequence

$$V_0 = 0, \quad V_{k+1} = F(U_k, V_k), \tag{3.3}$$

where $F : \mathbb{R}^{N_1+N_2} \to \mathbb{R}^{N_2}$ is a function that satisfies the following Lipschitz condition:

$$\|F(u,v) - F(\tilde{u},\tilde{v})\|_{\max} \leq L\|(u,v) - (\tilde{u},\tilde{v})\|_{\max} = L\max\{\|u-\tilde{u}\|_{\max}, \|v-\tilde{v}\|_{\max}\}, \tag{3.4}$$

where the Lipschitz constant $L$ is defined in Assumption 2.10.

Finally, we let $(\eta_t)_{t\geq 0}$ be an iid sequence of $\mathbb{R}$-valued random variables independent with $(U_t)_{t\geq 0}$ with distribution $\mu_\eta$ (hence also independent with $(U_t, V_t)_{t\geq 0}$). Then the joint process $(U_t, V_t, \eta_t)_{t\geq 0}$ in $\mathbb{R}^{N_1+N_2+1}$ is a Markov chain with transition kernel

$$P((u,v,\eta), A\times B\times C) = \mu_\eta(C)\delta_{F(u,v)}(B)\,\wp(u,A). \tag{3.5}$$

Under the above settings,

**Proposition 3.1.** *The transition kernel $P$ satisfies the contraction condition*

$$\frac{\mathsf{Wass}_2(P((u,v,\eta),\cdot), P((u',v',\eta'),\cdot))}{\|(u,v,\eta)-(u',v',\eta')\|_{\max}} \leq \sqrt{L^2 + L_\wp^2}. \tag{3.6}$$

*Proof.* By e.g. [17, Theorem 4.1], one could select optimal coupling $\gamma^*_{u,u'}$ between $\wp(u,\cdot)$ and $\wp(\tilde{u},\cdot)$. Defining $\Delta$ as the mapping $\eta \in \mathbb{R} \mapsto (\eta,\eta) \in \mathbb{R}^2$. Then

$$\tilde{\gamma}_{(u,v,\eta),(u',v',\eta')} := \gamma^*_{u,u'}\delta_{F(u,v)}\delta_{F(u',v')}(\Delta\#\mu_\eta)$$

is a coupling between $P((u,v),\cdot)$ and $P((u',v'),\cdot)$. Therefore,

$$\begin{aligned}
&\left(\mathsf{Wass}_2(P((u,v,\eta),\cdot),P((u',v',\eta'),\cdot))\right)^2\\
&\le \int_{\mathbb{R}^{N_1+N_2}} \|(\mathsf{u},\mathsf{v},\tilde{\eta}) - (\mathsf{u}',\mathsf{v}',\tilde{\eta}')\|^2_{\max}\,\tilde{\gamma}_{(u,v,\eta),(u',v',\eta')}(d\mathsf{u},d\mathsf{v},d\mathsf{u}',d\mathsf{v}',d\tilde{\eta},d\tilde{\eta}')\\
&= \int_{\mathbb{R}^{N_1}} \|(\mathsf{u},F(u,v),\tilde{\eta}) - (\mathsf{u}',F(u',v'),\tilde{\eta})\|^2_{\max}\,\gamma^*_{u,u'}(d\mathsf{u},d\mathsf{u}')\mu_\eta(d\eta)\\
&= \int_{\mathbb{R}^{N_1}} \left(\max\left(\|\mathsf{u}-\mathsf{u}'\|_{\max},\,\|F(u,v)-F(u',v')\|_{\max}\right)\right)^2\,\gamma^*_{u,u'}(d\mathsf{u},d\mathsf{u}')\\
&= \int_{\mathbb{R}^{N_1}} \max\left(\|\mathsf{u}-\mathsf{u}'\|^2_{\max},\,\|F(u,v)-F(u',v')\|^2_{\max}\right)\,\gamma^*_{u,u'}(d\mathsf{u},d\mathsf{u}')\\
&\le \int_{\mathbb{R}^{N_1}} \left(\|\mathsf{u}-\mathsf{u}'\|^2_{\max} + \|F(u,v)-F(u',v')\|^2_{\max}\right)\,\gamma^*_{u,u'}(d\mathsf{u},d\mathsf{u}')\\
&\le (L^2_\wp + L^2)\max(\|u-u'\|^2_{\max},\|v-v'\|^2_{\max})\\
&\le (L^2_\wp + L^2)\max(\|u-u'\|^2_{\max},\|v-v'\|^2_{\max},\|\eta-\eta'\|^2_{\max}).
\end{aligned}$$

□

We note that when $L^2 + L^2_\wp < 1$, then $Q$ is also a contraction. As a result, we can follow the arguments from remark 2.5 to show that -

**Theorem 3.2** (Existence of Unique Invariant Measure, Informal)**.** *The distribution of the joint Markov chain $(U_t,V_t)$, where $(V_t)$ is as defined in* (3.3) *with the associated $F$ satisfying the Lipschitzness condition* (3.4)*, converges in Wasserstein distance to an invariant measure $\rho^*$ such that $P^\vee\rho^* = \rho^*$, where*

$$P^\vee\rho(A) = \int P(\mathsf{h},A)\,d\mathsf{h}, \quad \mathsf{h} = (\mathsf{u},\mathsf{v},\tilde{\eta}). \tag{3.7}$$

Proposition 3.1 is crucial in developing the fixed point analysis for the joint process of the input sequence, hidden states, and (derivatives) of the memory layer. The evolution could be written as

$$H^{\theta,h}_t = (U_t, V^{\theta,h}_t, \eta_t), \quad U_t = (X_t,Z_t), \quad V^{\theta,h}_t = (S_{t,\theta,h},\tilde{S}_{t,\theta,h},\tilde{\tilde{S}}_{t,\theta,h}), \quad H^{\theta,h}_0 = h. \tag{3.8}$$

where $F := F^\theta$ is a deterministic function depending on both $U_t$ and $V_t$, which could be computed by the chain rule in the next section. If we are able to show that $F$ is Lipschitz in the sense of equation (3.4) (uniform in $\theta$), then we can apply proposition 3.1 to show that $H_t$ converges in distribution.

*Remark* 3.3. It is important to note that $F := F^\theta$, hence also the process $V^{\theta,h}$, all depend on the parameters of the RNN $\theta$. Note also that the process $H$ also depends the initial state $H_0 = h = (u,v,\eta)$, and for RNN $v = 0$ supposedly. We will, however, not impose this assumption for the convenience of carrying out the Poisson equation analysis in Chapter 5. Finally, we shall suppress the dependence of $\theta$ and $h$ when the context is clear.

### 3.2 The Lipschitzness of $F$

Before we proceed, we recall a few notations from the previous section

$$\tilde{S}_t^A = \frac{\partial S_t}{\partial A}, \quad \tilde{S}_t^W = \frac{\partial S_t}{\partial W}, \quad \tilde{S}_t^{A^{ij},k} = \frac{\partial S_{t+1}^k}{\partial A^{ij}}, \quad \tilde{S}_t^{W^{ij},k} = \frac{\partial S_t^k}{\partial W^{ij}}.$$

We further adopt the following notations: firstly, for the second derivatives

$$\begin{aligned}
\tilde{\tilde{S}}_t^{\theta_1,\theta_2} &= \frac{\partial S_t}{\partial \theta_1 \partial \theta_2}, \quad \tilde{\tilde{S}}_t^{AA} = \nabla_{AA}^2 S_t, \quad \tilde{\tilde{S}}_t^{AW} = \nabla_{AW}^2 S_t, \quad \tilde{\tilde{S}}_t^{WW} = \nabla_{WW}^2 S_t, \\
\tilde{\tilde{S}}_t^{A^{mn},A^{ij},:} &= \frac{\partial S_t}{\partial A^{mn} \partial A^{ij}}, \quad \tilde{\tilde{S}}_{t+1}^{A^{mn},W^{ij},:} = \frac{\partial S_t}{\partial A^{mn} \partial W^{ij}}, \quad \tilde{\tilde{S}}_{t+1}^{W^{mn},W^{ij},:} = \frac{\partial S_t}{\partial W^{mn} \partial W^{ij}}, \\
\tilde{\tilde{S}}_t^{A^{mn},A^{ij},k} &= \frac{\partial S_t^k}{\partial A^{mn} \partial A^{ij}}, \quad \tilde{\tilde{S}}_{t+1}^{A^{mn},W^{ij},k} = \frac{\partial S_t^k}{\partial A^{mn} \partial W^{ij}}, \quad \tilde{\tilde{S}}_{t+1}^{W^{mn},W^{ij},k} = \frac{\partial S_t^k}{\partial W^{mn} \partial W^{ij}}.
\end{aligned}$$

Let us also define the following dummy arguments for the input of function $F(u, v)$:

$$u = (x, z), \quad v = (s, \tilde{s}, \tilde{\tilde{s}}), \quad u' = (x', z'), \quad v' = (s', \tilde{s}', \tilde{\tilde{s}}'),$$

and adopt the index notation for $u, v$ (and similarly $u', v'$) as we have used to define the forward first/second derivatives:

$$\tilde{s}^{A_{ij},k}, \ \tilde{s}^{W_{ij},k}, \tilde{s}^{A_{mn}A_{ij},k}, \ \tilde{s}^{W_{mn}A_{ij},k} \text{ and } \tilde{s}^{W_{mn}W_{ij},k}.$$

Finally, we index the components of function $F := F^\theta$ so that

$$\begin{aligned}
S_{t+1}^k &= F_{S,k}^\theta(U_t, S_t) \overset{(*)}{=}: F_{S,k}^\theta(U_t, V_t), \\
\tilde{S}_{t+1}^{A^{ij},k} &= F_{A^{ij},k}^\theta(U_t, S_t, \tilde{S}_t) \overset{(*)}{=}: F_{A^{ij},k}^\theta(U_t, V_t), \\
\tilde{S}_{t+1}^{W^{ij},k} &= F_{W^{ij},k}^\theta(U_t, S_t, \tilde{S}_t) \overset{(*)}{=}: F_{W^{ij},k}^\theta(U_t, V_t), \\
\tilde{\tilde{S}}_{t+1}^{A^{mn},A^{ij},k} &= F_{A^{mn},A^{ij},k}^\theta(U_t, S_t, \tilde{S}_t, \tilde{\tilde{S}}_t) = F_{A^{mn},A^{ij},k}^\theta(U_t, V_t), \\
\tilde{\tilde{S}}_{t+1}^{A^{mn},W^{ij},k} &= F_{A^{mn},W^{ij},k}^\theta(U_t, S_t, \tilde{S}_t, \tilde{\tilde{S}}_t) = F_{A^{mn},W^{ij},k}^\theta(U_t, V_t), \\
\tilde{\tilde{S}}_{t+1}^{W^{mn},W^{ij},k} &= F_{W^{mn},W^{ij},k}^\theta(U_t, S_t, \tilde{S}_t, \tilde{\tilde{S}}_t) = F_{W^{mn},W^{ij},k}^\theta(U_t, V_t).
\end{aligned}$$

Note that with the above notations, we have

$$\frac{\partial F_S^\theta}{\partial A} = F_A^\theta, \quad \frac{\partial F_S^\theta}{\partial W} = F_W^\theta, \quad \frac{\partial F_A^\theta}{\partial A} = F_{AA}^\theta, \quad \frac{\partial F_W^\theta}{\partial A} = F_{AW}^\theta, \quad \frac{\partial F_W^\theta}{\partial W} = F_{WW}^\theta. \tag{3.9}$$

*Remark* 3.4. Here we include redundant dependence on $\tilde{S}_t$ and $\tilde{\tilde{S}}_t$ in equalities marked with $(*)$ as a notational convenience. We will also supress the dependence of $F := F^\theta$ on $\theta$ when there is no ambiguity.

Our objective is to prove that $F$ is Lipschitz. It is useful to note that

$$\begin{aligned}\|F(u,v)-F(\tilde{u},\tilde{v})\|_{\max} = &\max_k |F_{S,k}(u,v)-F_{S,k}(u,v)| \\ &\vee \max_{ijk} |F_{A^{ij},k}(u,v)-F_{A^{ij},k}(u,v)| \\ &\vee \max_{ijk} |F_{W^{ij},k}(u,v)-F_{W^{ij},k}(u,v)| \\ &\vee \max_{ijmnk} |F_{A^{mn},A^{ij},k}(u,v)-F_{A^{mn},A^{ij},k}(u,v)| \\ &\vee \max_{ijmnk} |F_{A^{mn}W^{ij}}(u,v)-F_{A^{mn},W^{ij},k}(u,v)| \\ &\vee \max_{ijmnk} |F_{W^{mn},W^{ij},k}(u,v)-F_{W^{mn},W^{ij},k}(u,v)|. \end{aligned} \tag{3.10}$$

Therefore to establish Lipschitzness with respect to the max norm, it suffices to control the increments of the individual components.

### 3.2.1 Lipschitzness of $F_{S,k}$

We take note of the explicit formulae of the components of $F$. Firstly, we have

$$F_{S,k}(u,v) = \sigma\left(\frac{1}{d}\left\langle \phi(A^{k,:}), x\right\rangle + \frac{1}{N}\left\langle \phi(W^{k,:}), s\right\rangle\right),$$

therefore

$$\begin{aligned}|F_{S,k}(u,v)-F_{S,k}(u',v')| &\le \frac{1}{4}\left|\frac{1}{d}\left\langle \phi(A^{k,:}), x\right\rangle + \frac{1}{N}\left\langle \phi(W^{k,:}), s\right\rangle - \frac{1}{d}\left\langle \phi(A^{k,:}), x'\right\rangle - \frac{1}{N}\left\langle \phi(W^{k,:}), s'\right\rangle\right| \\ &= \frac{1}{4}\left|\frac{1}{d}\sum_{\ell=1}^{d}\phi(A^{k,\ell})(x^\ell-(x')^\ell) + \frac{1}{N}\sum_{\ell=1}^{N}\phi(W^{k,\ell})(s^\ell-(s')^\ell)\right| \\ &\le \frac{1}{4}\left[\frac{1}{d}\left|\sum_{\ell=1}^{d}\phi(A^{k,\ell})(x^\ell-(x')^\ell)\right| + \frac{1}{N}\left|\sum_{\ell=1}^{N}\phi(W^{k,\ell})(s^\ell-(s')^\ell)\right|\right] \\ &\le \frac{1}{4}\left[\max_\ell\left|\phi(A^{k,\ell})\right|\left|(x^\ell-(x')^\ell)\right| + \max_\ell\left|\phi(W^{k,\ell})\right|\left|s^\ell-(s')^\ell)\right|\right] \\ &\le \frac{C_\phi}{4}\|x-x'\|_{\max} + \frac{C_\phi}{4}\|s-s'\|_{\max} \\ &\le \frac{C_\phi}{2}\max\left[\|x-x'\|_{\max}, \|s-s'\|_{\max}\right]. \end{aligned} \tag{3.11}$$

### 3.2.2 Explicit computations of $F$ and a-priori bounds

More work will be needed to study the other components. Firstly, we take note of:

$$\begin{aligned}F_{A^{ij},k}(u,v) &= \sigma'\left(\frac{1}{d}\left\langle \phi(A^{k,:}), x\right\rangle + \frac{1}{N}\left\langle \phi(W^{k,:}), s\right\rangle\right)\times\left(\frac{1}{d}\delta_{ik}\phi'(A^{ij})x^j + \frac{1}{N}\sum_{\ell=1}^{N}\phi(W^{k,\ell})\tilde{s}^{A^{ij},\ell}\right), \\ F_{W^{ij},k}(u,v) &= \sigma'\left(\frac{1}{d}\left\langle \phi(A^{k,:}), x\right\rangle + \frac{1}{N}\left\langle \phi(W^{k,:}), s\right\rangle\right)\times\left(\frac{1}{N}\delta_{ik}\phi'(W^{ij})s^j + \frac{1}{N}\sum_{\ell=1}^{N}\phi(W^{k\ell})\tilde{s}^{W^{ij},\ell}\right), \end{aligned} \tag{3.12}$$

Furthermore, we have

$$
\begin{aligned}
F_{A^{nm},A^{ij},k}(u,v) = \sigma''&\left(\frac{1}{d}\left\langle\phi(A^{k,:}),x\right\rangle + \frac{1}{N}\left\langle\phi(W^{k,:}),s\right\rangle\right)\\
&\times\left(\frac{1}{d}\delta_{ik}\phi'(A^{ij})x^j + \frac{1}{N}\sum_{\ell=1}^{N}\phi(W^{k\ell})\tilde{s}^{A^{ij},\ell}\right)\\
&\times\left(\frac{1}{d}\delta_{nk}\phi'(A^{nm})x^m + \frac{1}{N}\sum_{\ell=1}^{N}\phi(W^{k\ell})\tilde{s}^{A^{nm},\ell}\right)\\
&+\sigma'\left(\frac{1}{d}\left\langle\phi(A^{k,:}),x\right\rangle + \frac{1}{N}\left\langle\phi(W^{k,:}),s\right\rangle\right)\\
&\times\left(\frac{1}{d}\delta_{ik}\delta_{ij=nm}\phi''(A^{ij})x^j + \frac{1}{N}\sum_{\ell=1}^{N}\phi(W^{k\ell})\tilde{\tilde{s}}^{A^{ij},A^{nm},\ell}\right). \qquad (3.13)\\
F_{W^{nm},A^{ij},k}(u,v) = \sigma''&\left(\frac{1}{d}\left\langle\phi(A^{k,:}),x\right\rangle + \frac{1}{N}\left\langle\phi(W^{k,:}),s\right\rangle\right)\\
&\times\left(\frac{1}{N}\delta_{kn}\phi'(W^{nm})s^m + \frac{1}{N}\sum_{\ell=1}^{N}\phi(W^{k\ell})\tilde{s}^{W^{nm},\ell}\right)\\
&\times\left(\frac{1}{d}\delta_{ik}\phi'(A^{ij})x^j + \frac{1}{N}\sum_{\ell=1}^{N}\phi(W^{k\ell})\tilde{s}^{A^{ij},\ell}\right)\\
&+\sigma'\left(\frac{1}{d}\left\langle\phi(A^{k,:}),x\right\rangle + \frac{1}{N}\left\langle\phi(W^{k,:}),s\right\rangle\right)\\
&\times\left(\frac{1}{N}\delta_{kn}\phi'(W^{nm})\tilde{s}^{A^{ij},m} + \frac{1}{N}\sum_{\ell=1}^{N}\phi(W^{k\ell})\tilde{\tilde{s}}^{W^{nm},A^{ij},\ell}\right). \qquad (3.14)\\
F_{W^{nm},W^{ij},k}(u,v) = \sigma''&\left(\frac{1}{d}\left\langle\phi(A^{k,:}),x\right\rangle + \frac{1}{N}\left\langle\phi(W^{k,:}),s\right\rangle\right)\\
&\times\left(\frac{1}{N}\delta_{kn}\phi'(W^{nm})s^m + \frac{1}{N}\sum_{\ell=1}^{N}\phi(W^{k\ell})\tilde{s}^{W^{nm},\ell}\right)\\
&\times\left(\frac{1}{N}\delta_{ik}\phi'(W^{ij})s^j + \frac{1}{N}\sum_{\ell=1}^{N}\phi(W^{k\ell})\tilde{s}^{W^{ij},\ell}\right)\\
&+\sigma'\left(\frac{1}{d}\left\langle\phi(A^{k\ell}),x\right\rangle + \frac{1}{N}\left\langle\phi(W^{k,:}),s\right\rangle\right)\\
&\times\left(\frac{1}{N}\delta_{ik}\delta_{ij=nm}\phi''(W^{ij})s^j + \frac{1}{N}\sum_{\ell=1}^{N}\phi(W^{k\ell})\tilde{\tilde{s}}^{W^{nm},W^{ij},\ell}\right). \qquad (3.15)
\end{aligned}
$$

### 3.2.3 Bounds for Gradient Process

Let's establish an upper bound for the forward gradient processes.

**Lemma 3.5.**

$$
\begin{aligned}
\|\tilde{S}_0^A\|_{\max}\vee\|\hat{S}_0^A\|_{\max} \le \frac{C_{\phi'}}{4-C_\phi}\frac{1}{d} \implies \|\tilde{S}_t^A\|_{\max}\vee\|\hat{S}_t^A\|_{\max} \le \frac{C_{\phi'}}{4-C_\phi}\frac{1}{d},\\
\|\tilde{S}_0^W\|_{\max}\vee\|\hat{S}_0^W\|_{\max} \le \frac{C_{\phi'}}{4-C_\phi}\frac{1}{N}. \implies \|\tilde{S}_t^W\|_{\max}\vee\|\hat{S}_t^W\|_{\max} \le \frac{C_{\phi'}}{4-C_\phi}\frac{1}{N}. \qquad (3.16)
\end{aligned}
$$

*Proof.* We begin by noting that

$$\begin{aligned}|\tilde{S}_{t+1}^{A^{ij},k}| &= |F_{A^{ij},k}(U_t, V_t)| \\ &\le \max|\sigma'(\cdot)|\left(\frac{1}{d}\delta_{ik}|\phi'(A^{ij})X_t^j| + \frac{1}{N}\left|\sum_{\ell=1}^{N}\phi(W^{k\ell})\tilde{S}_t^{A^{ij},\ell}\right|\right) \\ &\le \frac{1}{4}\left(\frac{C_{\phi'}}{d} + C_\phi \max_\ell |\tilde{S}_t^{A^{ij},\ell}|\right), \end{aligned} \tag{3.17}$$

where in the last inequality, we used the upper bound for $\sigma'(y)$ by Lemma 2.9. This bound then implies that

$$\max_k |\tilde{S}_{t+1}^{A^{ij},k}| \le \frac{C_{\phi'}}{4d} + \frac{C_\phi}{4}\max_k |\tilde{S}_t^{A^{ij},k}|. \tag{3.18}$$

Note that $\tilde{S}_0^{A^{ij},k} = 0$, so by recursive inequality (B.5), for $C_\phi < 4$,

$$\max_\ell |\tilde{S}_{t+1}^{A^{ij},\ell}| \le \left(\frac{C_\phi}{4}\right)^k \frac{C_{\phi'}}{(4-C_\phi)d} + \frac{1-(C_\phi/4)^k}{1-C_\phi/4}\frac{C_{\phi'}}{4d} = \frac{C_{\phi'}}{d(4-C_\phi)} \quad \text{for any } t \ge 0. \tag{3.19}$$

Similarly, we have

$$\begin{aligned}|\tilde{S}_{t+1}^{W^{ij},k}| &= |F_{W^{ij},k}(U_t, V_t)| \\ &\le \frac{1}{4}\left(\frac{1}{N}\delta_{ik}|\phi'(W^{ij})S_t^j| + \frac{1}{N}\left|\sum_{\ell=1}^{N}\phi(W^{k\ell})\tilde{S}_t^{W^{ij},\ell}\right|\right) \\ &\le \frac{1}{4}\left(\frac{C_{\phi'}}{N} + C_\phi \max_k |\tilde{S}_t^{W^{ij},k}|\right), \end{aligned} \tag{3.20}$$

which implies

$$\max_k |\tilde{S}_{t+1}^{W^{ij},k}| \le \frac{C_{\phi'}}{4N} + \frac{C_\phi}{4}\max_k |\tilde{S}_t^{W^{ij},k}|. \tag{3.21}$$

Since $\tilde{S}_0^{W^{ij},k} = 0$, so by recursive inequality (B.5), for $C_\phi < 4$,

$$\max_k |\tilde{S}_{t+1}^{W^{ij},k}| \le \frac{C_{\phi'}}{N(4-C_\phi)}. \tag{3.22}$$

Using exactly the same calculations as above, we can show the bounds for $\hat{S}_t^A$ and $\hat{S}_t^W$. □

*Remark* 3.6. As the proof of Lemma 3.5 shows the denominator of the upper bounds for the forward gradient processes presented in Lemma 3.5 has to do with the inverse of the upper bound of the derivative of the sigmoid activation function by Lemma 2.9. That is, the proofs can accommodate other activation functions too as long as they are bounded with bounded derivatives and sufficiently smooth.

In a similar fashion, we can also establish bound for the second-order derivative process $\tilde{\tilde{S}}_t$.

**Lemma 3.7.** *Assume that Assumptions 2.3 and 2.10 hold, then there are uniform constants $a_{AA,\max}$, $a_{WA,\max}$ and $a_{WW,\max}$ independent of $t \ge 0$, such that the following implications hold:*

$$\begin{aligned}\|\tilde{\tilde{S}}_0^{AA}\|_{\max} \le \frac{a_{AA,\max}}{1-(C_\phi/4)} &\implies \|\tilde{\tilde{S}}_t^{AA}\|_{\max} \le \frac{a_{AA,\max}}{1-(C_\phi/4)} \\ \|\tilde{\tilde{S}}_0^{WA}\|_{\max} \le \frac{a_{WA,\max}}{1-(C_\phi/4)} &\implies \|\tilde{\tilde{S}}_t^{WA}\|_{\max} \le \frac{a_{WA,\max}}{1-(C_\phi/4)} \\ \|\tilde{\tilde{S}}_0^{WW}\|_{\max} \le \frac{a_{WW,\max}}{1-(C_\phi/4)} &\implies \|\tilde{\tilde{S}}_t^{WW}\|_{\max} \le \frac{a_{WW,\max}}{1-(C_\phi/4)} \end{aligned} \tag{3.23}$$

*Proof.* We begin by noting that

$$\begin{aligned}
|\tilde{\tilde{S}}_{t+1}^{A^{nm},A^{ij},k}| &= |F_{A^{nm},A^{ij},k}(U_t,V_t)| \\
&\le \frac{1}{10}\left|\frac{1}{d}\delta_{ik}\phi'(A^{ij})X_t^j + \frac{1}{N}\sum_{\ell=1}^{N}\phi(W^{k\ell})\tilde{S}_t^{A^{ij},\ell}\right| \\
&\quad\times\left|\frac{1}{d}\delta_{nk}\phi'(A^{nm})X_t^m + \frac{1}{N}\sum_{\ell=1}^{N}\phi(W^{k\ell})\tilde{S}_t^{A^{nm},\ell}\right| \\
&\quad+\frac{1}{4}\left|\frac{1}{d}\delta_{ik}\delta_{ij=nm}\phi''(A^{ij})X_t^j + \frac{1}{N}\sum_{\ell=1}^{N}\phi(W^{k\ell})\tilde{\tilde{S}}_t^{A^{ij},A^{nm},\ell}\right| \\
&\le \frac{1}{10}\left(\frac{C_{\phi'}}{d} + C_\phi\max_k|\tilde{S}_t^{A^{ij},k}|\right)\left(\frac{C_{\phi'}}{d} + C_\phi\max_k|\tilde{S}_t^{A^{nm},k}|\right) \\
&\quad+\frac{1}{4}\left(\frac{C_{\phi''}}{d} + C_\phi\max_k|\tilde{\tilde{S}}_t^{A^{ij},A^{nm},k}|\right) \\
&\le a_{AA,\max} + \frac{C_\phi}{4}\max_k|\tilde{\tilde{S}}_t^{A^{ij},A^{nm},k}|, \qquad (3.24)
\end{aligned}$$

where, recalling that $M_\phi = 1 + [C_\phi/(4 - C_\phi)]$,

$$\begin{aligned}
a_{AA,\max} &= \frac{1}{10}\left(\frac{C_{\phi'}}{d} + C_\phi\frac{C_{\phi'}}{d(4-C_\phi)}\right)\left(\frac{C_{\phi'}}{d} + C_\phi\frac{C_{\phi'}}{d(4-C_\phi)}\right) + \frac{1}{4}\frac{C_{\phi''}}{d} \\
&= \frac{C_{\phi'}^2M_\phi^2}{10d^2} + \frac{C_{\phi''}}{4d}. \qquad (3.25)
\end{aligned}$$

Therefore, we have

$$\max_k|\tilde{\tilde{S}}_{t+1}^{A^{ij},A^{nm},k}| \le a_{AA,\max} + \frac{C_\phi}{4}\max_k|\tilde{\tilde{S}}_t^{A^{ij},A^{nm},k}|. \qquad (3.26)$$

Note that $\tilde{\tilde{S}}_0^{A^{ij},A^{nm},k} = 0$, so by recursive inequality (B.1), for $C_\phi < 4$,

$$\max_k|\tilde{\tilde{S}}_{t+1}^{A^{ij},A^{nm},k}| \le \frac{a_{AA,\max}}{1 - (C_\phi/4)}. \qquad (3.27)$$

Next, we have

$$\begin{aligned}
|\tilde{\tilde{S}}_{t+1}^{W^{nm},A^{ij},k}| &= |F_{W^{nm},A^{ij},k}(U_t,V_t)| \\
&\le \frac{1}{10}\left|\frac{1}{N}\delta_{kn}\phi'(W^{nm})S_t^m + \frac{1}{N}\sum_{\ell=1}^{N}\phi(W^{k\ell})\tilde{S}_t^{W^{nm},\ell}\right| \\
&\quad\times\left|\frac{1}{d}\delta_{ik}\phi'(A^{ij})X_t^j + \frac{1}{N}\sum_{\ell=1}^{N}\phi(W^{k\ell})\tilde{S}_t^{A^{ij},\ell}\right| \\
&\quad+\frac{1}{4}\left|\frac{1}{N}\delta_{kn}\phi'(W^{nm})\tilde{S}_t^{A^{ij},m} + \frac{1}{N}\phi(W^{k\ell})\tilde{\tilde{S}}_t^{W^{nm},A^{ij},\ell}\right| \\
&\le \frac{1}{10}\left(\frac{C_{\phi'}}{N} + C_\phi\max_k|\tilde{S}_t^{W^{nm},k}|\right)\left(\frac{C_{\phi'}}{d} + C_\phi\max_k|\tilde{S}_t^{A^{ij},k}|\right) \\
&\quad+\frac{1}{4}\left(\frac{C_{\phi'}}{N}\max_k|\tilde{S}^{A^{ij},k}| + C_\phi\max_k|\tilde{\tilde{S}}_t^{W^{nm},A^{ij},k}|\right) \\
&\le a_{WA,\max} + \frac{C_\phi}{4}\max_k|\tilde{\tilde{S}}_t^{W^{nm},A^{ij},k}|, \qquad (3.28)
\end{aligned}$$

where

$$a_{WA,\max} = \frac{1}{10}\left(\frac{C_{\phi'}}{N} + C_\phi \frac{C_{\phi'}}{N(4-C_\phi)}\right)\left(\frac{C_{\phi'}}{d} + C_\phi \frac{C_{\phi'}}{d(4-C_\phi)}\right) + \frac{1}{4}\frac{C_{\phi'}}{N}\frac{C_{\phi'}}{d(4-C_\phi)}$$
$$= \frac{C_{\phi'}^2 M_\phi^2}{10Nd} + \frac{C_{\phi'}^2}{4Nd(4-C_\phi)}. \tag{3.29}$$

Since $\tilde{\tilde{S}}_0^{W^{nm},A^{ij},k} = 0$, so by recursive inequality (B.1), for $C_\phi < 4$,

$$\max_k |\tilde{\tilde{S}}_{t+1}^{W^{nm},A^{ij},k}| \le \frac{a_{WA,\max}}{1-(C_\phi/4)}. \tag{3.30}$$

Finally,

$$\begin{aligned}
|\tilde{\tilde{S}}_{t+1}^{W^{nm},W^{ij},k}| &= |F_{W^{nm},W^{ij},k}(U_t, V_t)| \\
&= \frac{1}{10}\left|\frac{1}{N}\delta_{kn}\phi'(W^{nm})S_t^m + \frac{1}{N}\sum_{\ell=1}^N \phi(W^{k\ell})\tilde{S}_t^{W^{nm},\ell}\right| \\
&\quad \times \left|\frac{1}{N}\delta_{ik}\phi'(W^{ij})S_t^j + \frac{1}{N}\sum_{\ell=1}^N \phi(W^{k\ell})\tilde{S}_t^{W^{ij},\ell}\right| \\
&\quad + \frac{1}{4} \times \left|\frac{1}{N}\delta_{ik}\delta_{ij=nm}\phi''(W^{ij})S_t^j + \frac{1}{N}\sum_{\ell=1}^N \phi(W^{k\ell})\tilde{\tilde{S}}_t^{W^{nm},W^{ij},\ell}\right| \\
&\le \frac{1}{10}\left(\frac{C_{\phi'}}{N} + C_\phi \max_k \left|\tilde{S}_t^{W^{nm},k}\right|\right)\left(\frac{C_{\phi'}}{N} + C_\phi \max_k \left|\tilde{S}_t^{W^{ij},k}\right|\right) \\
&\quad + \frac{1}{4}\left(\frac{C_{\phi'}}{N} + C_\phi \max_k |\tilde{\tilde{S}}_t^{W^{nm},W^{ij},k}|\right) \\
&\le a_{WW,\max} + \frac{C_\phi}{4}\max_k |\tilde{\tilde{S}}_t^{W^{nm},W^{ij},k}|,
\end{aligned} \tag{3.31}$$

where

$$a_{WW,\max} = \frac{1}{10}\left(\frac{C_{\phi'}}{N} + C_\phi \frac{C_{\phi'}}{N(4-C_\phi)}\right)\left(\frac{C_{\phi'}}{N} + C_\phi \frac{C_{\phi'}}{N(4-C_\phi)}\right) + \frac{C_{\phi''}}{4N} = \frac{C_{\phi'}^2 M_\phi^2}{10N^2} + \frac{C_{\phi''}}{4N}. \tag{3.32}$$

Again $\tilde{\tilde{S}}_0^{W^{nm},A^{ij},k} = 0$, so by recursive inequality (B.1), for $C_\phi < 4$,

$$\max_k |\tilde{\tilde{S}}_{t+1}^{W^{nm},W^{ij},k}| \le \frac{a_{WW,\max}}{1-(C_\phi/4)}. \tag{3.33}$$

□

From the above bounds, we conclude that the process $H_t$ takes its value in the hyper-rectangle

$$\begin{aligned}
\mathcal{R} = [-1,1]^{d+d_Z} \times [0,1]^N \times \left[-\frac{C_{\phi'}}{d(4-C_\phi)}\right]^{N\times Nd} \times \left[-\frac{C_{\phi'}}{N(4-C_\phi)}\right]^{N\times N^2} \\
\times \left[-\frac{a_{AA,\max}}{1-(C_\phi/4)}, \frac{a_{AA,\max}}{1-(C_\phi/4)}\right]^{N\times(Nd)(Nd)} \times \left[-\frac{a_{WA,\max}}{1-(C_\phi/4)}, \frac{a_{WA,\max}}{1-(C_\phi/4)}\right]^{N\times(N^2)(Nd)} \\
\times \left[-\frac{a_{WW,\max}}{1-(C_\phi/4)}, \frac{a_{WW,\max}}{1-(C_\phi/4)}\right]^{N\times(N^2)(N^2)}.
\end{aligned}$$

The boundedness of process $H_t$ is essential for the Lipschitzness of $F$.

### 3.2.4 Lipschitzness of $F$

We may now prove that

**Lemma 3.8.** *If $(u,v),(u',v') \in \mathcal{R}$, then there exists $L_0 > 0$ (as defined in Assumption 2.10) such that*

$$\begin{aligned}\max_{ijk} |F_{A^{ij},k}(u,v) - F_{A^{ij},k}(u',v')| &\le L_0 \max\left[\|u-u'\|_{\max}, \|v-v'\|_{\max}\right], \\ \max_{ijk} |F_{W^{ij},k}(u,v) - F_{W^{ij},k}(u',v')| &\le L_0 \max\left[\|u-u'\|_{\max}, \|v-v'\|_{\max}\right]. \end{aligned} \tag{3.34}$$

*Proof.* For simplicity, we denote $M_\phi = 1 + [C_\phi/(4-C_\phi)]$. Then

$$\begin{aligned}
&|F_{A^{ij},k}(u,v) - F_{A^{ij},k}(u',v')| \\
&\le \frac{1}{10} \times \left| \frac{1}{d}\delta_{ik}\phi'(A^{ij})x^j + \frac{1}{N}\sum_{\ell=1}^N \phi(W^{k\ell})\tilde{s}^{A^{ij},\ell} \right| \\
&\quad \times \left| \frac{1}{d}\left\langle \phi(A^{k,\cdot}), x - x' \right\rangle + \frac{1}{N}\left\langle \phi(W_t^{k,\cdot}), s - s' \right\rangle \right| \\
&\quad + \frac{1}{4} \times \left| \frac{1}{d}\delta_{ik}\phi'(A^{ij})(x^j - (x')^j) + \frac{1}{N}\sum_{\ell=1}^N \phi(W^{k\ell})(\tilde{s}^{A^{ij},\ell} - (\tilde{s}')^{A^{ij},\ell}) \right| \\
&\le \frac{1}{10}\left( \frac{C_{\phi'}}{d} + C_\phi \frac{C_{\phi'}}{d(4-C_\phi)} \right)\left( C_\phi \|x-x'\|_{\max} + C_\phi \|s-s'\|_{\max} \right) \\
&\quad + \frac{1}{4} \times \left( \frac{C_{\phi'}}{d}\|x-x'\|_{\max} + C_\phi \|\tilde{s}-\tilde{s}'\|_{\max} \right). \\
&\le \frac{C_{\phi'}}{d}\left( \frac{C_\phi M_\phi}{10} + \frac{1}{4} \right)\|x-x'\|_{\max} + \frac{C_\phi M_\phi C_{\phi'}}{10d}\|s-s'\|_{\max} + \frac{C_\phi}{4}\|\tilde{s}-\tilde{s}'\|_{\max} \\
&\le \max\left[ \frac{3C_{\phi'}}{d}\left( \frac{C_\phi M_\phi}{10} + \frac{1}{4} \right), \frac{3C_\phi}{4} \right] \times \max\left[ \|x-x'\|_{\max}, \|s-s'\|_{\max}, \|\tilde{s}-\tilde{s}'\|_{\max} \right].
\end{aligned} \tag{3.35}$$

Similarly,

$$\begin{aligned}
&|F_{W^{ij},k}(u,v) - F_{W^{ij},k}(u',v')| \\
&\le \frac{1}{10} \times \left| \frac{1}{N}\delta_{ik}\phi'(W^{ij})s^j + \frac{1}{N}\sum_{\ell=1}^N \phi(W^{k\ell})\tilde{s}^{W^{ij},\ell} \right| \\
&\quad \times \left| \frac{1}{d}\left\langle \phi(A^{k,\cdot}), x - x' \right\rangle + \frac{1}{N}\left\langle \phi(W^{k,\cdot}), s - s' \right\rangle \right| \\
&\quad + \frac{1}{4} \times \left| \frac{1}{N}\delta_{ik}\phi'(W^{ij})(s^j - (s')^j) + \frac{1}{N}\sum_{\ell=1}^N \phi(W^{k\ell}) \cdot (\tilde{s}^{W^{ij},\ell} - (\tilde{s}')^{W^{ij},\ell}) \right| \\
&\le \frac{1}{10} \times \left( \frac{C_{\phi'}}{N} + C_\phi \frac{C_{\phi'}}{N(4-C_\phi)} \right) \times \left( C_\phi \|x-x'\|_{\max} + C_\phi \|s-s'\|_{\max} \right) \\
&\quad + \frac{1}{4} \times \left( \frac{C_{\phi'}}{N}\|s-s'\|_{\max} + C_\phi \|\tilde{s}-\tilde{s}'\|_{\max} \right) \\
&\le \frac{C_\phi M_\phi C_{\phi'}}{10N}\|x-x'\|_{\max} + \frac{C_{\phi'}}{N}\left( \frac{C_\phi M_\phi}{10} + \frac{1}{4} \right)\|s-s'\|_{\max} + \frac{C_\phi}{4}\|\tilde{s}-\tilde{s}'\|_{\max} \\
&\le \max\left[ \frac{3C_{\phi'}}{N}\left( \frac{C_\phi M_\phi}{10} + \frac{1}{4} \right), \frac{3C_\phi}{4} \right] \times \max\left[ \|x-x'\|_{\max}, \|s-s'\|_{\max}, \|\tilde{s}-\tilde{s}'\|_{\max} \right].
\end{aligned} \tag{3.36}$$

Collecting the bounds above, we obtain:

$$\begin{aligned}
&\max_{ij}\left[|F_{A^{ij},k}(u,v)-F_{A^{ij},k}(u',v')|\vee|F_{W^{ij},k}(u,v)-F_{W^{ij},k}(u',v')|\right]\\
&\le \underbrace{\max\left(\frac{3C_{\phi'}}{\min(N,d)}\left(\frac{C_\phi M_\phi}{10}+\frac{1}{4}\right),\frac{3C_\phi}{4}\right)}_{:=L_0}\max\left[\|x-x'\|_{\max},\|s-s'\|_{\max},\|\tilde{s}-\tilde{s}'\|_{\max}\right].
\end{aligned}$$

□

Similarly, we have

**Lemma 3.9.** *If $(u,v),(u',v')\in\mathcal{R}$, then there exists $L_1>0$ (as defined in Assumption 2.10) such that*

$$\begin{aligned}
\max_{ijnmk}|F_{A^{nm},A^{ij},k}(u,v)-F_{A^{nm},A^{ij},k}(u',v')|&\le L_1\max\left[\|u-u'\|_{\max},\|v-v'\|_{\max}\right],\\
\max_{ijnmk}|F_{W^{nm},A^{ij},k}(u,v)-F_{W^{nm},A^{ij},k}(u',v')|&\le L_1\max\left[\|u-u'\|_{\max},\|v-v'\|_{\max}\right],\\
\max_{ijnmk}|F_{W^{nm},W^{ij},k}(u,v)-F_{W^{nm},A^{ij},k}(u',v')|&\le L_1\max\left[\|u-u'\|_{\max},\|v-v'\|_{\max}\right].
\end{aligned}\tag{3.37}$$

*Proof.* Define

$$\begin{aligned}
E_{\mathsf{I},k}&=\sigma''\left(\frac{1}{d}\left\langle\phi(A^{k,:}),x\right\rangle+\frac{1}{N}\left\langle\phi(W^{k,:}),s\right\rangle\right),\\
E_{\mathsf{II},k}&=\sigma'\left(\frac{1}{d}\left\langle\phi(A^{k,:}),x\right\rangle+\frac{1}{N}\left\langle\phi(W^{k,:}),s\right\rangle\right),\\
E_{\mathsf{III},ijk}&=\frac{1}{d}\delta_{ik}\phi'(A^{ij})x^j+\frac{1}{N}\sum_{\ell=1}^N\phi(W^{k\ell})\tilde{s}_1^{A^{ij},\ell},\\
E_{\mathsf{IV},ijk}&=\frac{1}{N}\delta_{ik}\phi'(W^{ij})s^j+\frac{1}{N}\sum_{\ell=1}^N\phi(W^{k\ell})\tilde{s}_1^{W^{ij},\ell},\\
E_{\mathsf{V},nmijk}&=\frac{1}{d}\delta_{ik}\delta_{ij=nm}\phi''(A^{ij})x^j+\frac{1}{N}\sum_{\ell=1}^N\phi(W^{k\ell})\tilde{\tilde{s}}^{A^{ij},A^{nm},\ell}\\
E_{\mathsf{VI},nmijk}&=\frac{1}{N}\delta_{kn}\phi'(W^{nm})\tilde{s}^{A^{ij},m}+\frac{1}{N}\sum_{\ell=1}^N\phi(W^{k\ell})\tilde{\tilde{s}}^{W^{nm},A^{ij},\ell}\\
E_{\mathsf{VII},nmijk}&=\frac{1}{N}\delta_{ik}\delta_{ij=nm}\phi''(W^{ij})s^j+\frac{1}{N}\sum_{\ell=1}^N\phi(W^{k\ell})\tilde{\tilde{s}}_1^{W^{nm},W^{ij},\ell},
\end{aligned}$$

so that

$$\begin{aligned}
F_{A^{nm},A^{ij},k}(u,v)&=E_{\mathsf{I},k}E_{\mathsf{III},nmk}E_{\mathsf{III},ijk}+E_{\mathsf{II},k}E_{\mathsf{V},nmijk}\\
F_{W^{nm},A^{ij},k}(u,v)&=E_{\mathsf{I},k}E_{\mathsf{IV},nmk}E_{\mathsf{III},ijk}+E_{\mathsf{II},k}E_{\mathsf{VI},nmijk}\\
F_{W^{nm},W^{ij},k}(u,v)&=E_{\mathsf{I},k}E_{\mathsf{IV},nmk}E_{\mathsf{IV},ijk}+E_{\mathsf{II},k}E_{\mathsf{VII},nmijk}
\end{aligned}$$

Similarly, define

$$
\begin{aligned}
E'_{\mathsf{I},k} &= \sigma''\left(\frac{1}{d}\left\langle\phi(A^{k,:}),x'\right\rangle + \frac{1}{N}\left\langle\phi(W^{k,:}),s'\right\rangle\right),\\
E'_{\mathsf{II},k} &= \sigma'\left(\frac{1}{d}\left\langle\phi(A^{k,:}),x'\right\rangle + \frac{1}{N}\left\langle\phi(W^{k,:}),s'\right\rangle\right),\\
E'_{\mathsf{III},ijk} &= \frac{1}{d}\delta_{ik}\phi'(A^{ij})(x')^j + \frac{1}{N}\sum_{\ell=1}^{N}\phi(W^{k\ell})(\tilde{s}')^{A^{ij},\ell},\\
E'_{\mathsf{IV},ijk} &= \frac{1}{N}\delta_{ik}\phi'(W^{ij})(s')^j + \frac{1}{N}\sum_{\ell=1}^{N}\phi(W^{k\ell})(\tilde{s}')^{W^{ij},\ell},\\
E'_{\mathsf{V},nmijk} &= \frac{1}{d}\delta_{ik}\delta_{ij=nm}\phi''(A^{ij})(x')^j + \frac{1}{N}\sum_{\ell=1}^{N}\phi(W^{k\ell})(\tilde{\tilde{s}}')^{A^{ij},A^{nm},\ell}\\
E'_{\mathsf{VI},nmijk} &= \frac{1}{N}\delta_{kn}\phi'(W^{nm})(\tilde{s}')^{A^{ij},m} + \frac{1}{N}\sum_{\ell=1}^{N}\phi(W^{k\ell})(\tilde{\tilde{s}}')^{W^{nm},A^{ij},\ell}\\
E'_{\mathsf{VII},nmijk} &= \frac{1}{N}\delta_{ik}\delta_{ij=nm}\phi''(W^{ij})(s')^j + \frac{1}{N}\sum_{\ell=1}^{N}\phi(W^{k\ell})(\tilde{\tilde{s}}')^{W^{nm},W^{ij},\ell}.
\end{aligned}
$$

Then,

$$
\begin{aligned}
F_{A^{nm},A^{ij},k}(u,v) - F_{A^{nm},A^{ij},k}(u',v') &= (E_{\mathsf{I},k}E_{\mathsf{III},nmk}E_{\mathsf{III},ijk} - E'_{\mathsf{I},k}E'_{\mathsf{III},nmk}E'_{\mathsf{III},ijk})\\
&\quad + (E_{\mathsf{II},k}E_{\mathsf{V},nmijk} - E'_{\mathsf{II},k}E'_{\mathsf{V},nmijk})\\
F_{W^{nm},A^{ij},k}(u,v) - F_{W^{nm},A^{ij},k}(u',v') &= (E_{\mathsf{I},k}E_{\mathsf{IV},nmk}E_{\mathsf{III},ijk} - E'_{\mathsf{I},k}E'_{\mathsf{IV},nmk}E'_{\mathsf{III},ijk})\\
&\quad + (E_{\mathsf{II},k}E_{\mathsf{VI},nmijk} - E'_{\mathsf{II},k}E'_{\mathsf{VI},nmijk})\\
F_{W^{nm},W^{ij},k}(u,v) - F_{W^{nm},A^{ij},k}(u',v') &= (E_{\mathsf{I},k}E_{\mathsf{IV},nmk}E_{\mathsf{IV},ijk} - E_{\mathsf{I},k}E_{\mathsf{IV},nmk}E_{\mathsf{IV},ijk})\\
&\quad + (E_{\mathsf{II},k}E_{\mathsf{VII},nmijk} - E'_{\mathsf{II},k}E'_{\mathsf{VII},nmijk})
\end{aligned}
$$

By our earlier calculations,

$$
\begin{gathered}
\max_k|E_{\mathsf{I},k}| \le \frac{1}{10},\quad \max_k|E_{\mathsf{II},k}| \le \frac{1}{4},\quad \max_k|E_{\mathsf{III},ijk}| \le \frac{C_{\phi'}M_\phi}{d},\quad \max_k|E_{\mathsf{IV},ijk}| \le \frac{C_{\phi'}M_\phi}{N},\\
\max_k|E_{\mathsf{V},nmijk}| \le \frac{1}{d}C_{\phi''} + \frac{C_\phi}{1-(C_\phi/4)}a_{AA,\max},\quad \max_k|E_{\mathsf{VI},nmijk}| \le \frac{1}{Nd}\frac{C_{\phi'}^2}{4-C_\phi} + \frac{C_\phi}{1-(C_\phi/4)}a_{WA,\max}\\
\max_k|E_{\mathsf{VII},nmijk}| \le \frac{1}{N}\delta_{ij=nm}C_{\phi''} + \frac{C_\phi}{1-(C_\phi/4)}a_{WW,\max}
\end{gathered}
$$

Furthermore, we compute the increments by suitable Taylor's expansion, noting that $\max_y |\sigma'''(y)| \le 1/8$:

$$
\begin{aligned}
|E_{\mathsf{I},k} - E'_{\mathsf{I},k}| &\le \frac{C_\phi}{8}\left(\|x - x'\|_{\max} + \|s - s'\|_{\max}\right)\\
|E_{\mathsf{II},k} - E'_{\mathsf{II},k}| &\le \frac{C_\phi}{10}(\|x - x'\|_{\max} + \|s - s'\|_{\max}),\\
|E_{\mathsf{III},ijk} - E'_{\mathsf{III},ijk}| &\le \frac{1}{d}C_{\phi'}|x^j - (x')^j| + C_\phi\|\tilde{s} - \tilde{s}'\|_{\max},\\
|E_{\mathsf{IV},ijk} - E'_{\mathsf{IV},ijk}| &\le \frac{1}{N}C_{\phi'}|s^j - (s')^j| + C_\phi\|\tilde{s} - \tilde{s}'\|_{\max},\\
|E_{\mathsf{V},nmijk} - E'_{\mathsf{V},nmijk}| &\le \frac{1}{d}C_{\phi''}|x^j - (x')^j| + C_\phi\|\tilde{\tilde{s}} - \tilde{\tilde{s}}'\|_{\max}\\
|E_{\mathsf{VI},nmijk} - E'_{\mathsf{VI},nmijk}| &\le \frac{1}{N}C_{\phi'}\|\tilde{s} - \tilde{s}'\|_{\max} + C_\phi\|\tilde{\tilde{s}} - \tilde{\tilde{s}}'\|_{\max}\\
|E_{\mathsf{VII},nmijk} - E'_{\mathsf{VII},nmijk}| &\le \frac{1}{N}C_{\phi''}|s^j - (s')^j| + C_\phi\|\tilde{\tilde{s}} - \tilde{\tilde{s}}'\|_{\max}.
\end{aligned}
$$

Collecting all of the above bounds via the inequality

$$|abc - a'b'c'| \le |ab||c - c'| + |ac'||b - b'| + |b'c'||a - a'|,$$

we have:

$$
\begin{aligned}
&\left|F_{A^{nm},A^{ij},k}(u,v) - F_{A^{nm},A^{ij},k}(u',v')\right|\\
&\le \left|E_{\mathsf{I},k}E_{\mathsf{III},nmk}\right|\left|E_{\mathsf{III},ijk} - E'_{\mathsf{III},ijk}\right| + \left|E_{\mathsf{I},k}E'_{\mathsf{III},ijk}\right|\left|E_{\mathsf{III},nmk} - E'_{\mathsf{III},nmk}\right| + \left|E'_{\mathsf{III},nmk}E'_{\mathsf{III},ijk}\right|\left|E_{\mathsf{I},k} - E'_{\mathsf{I},k}\right|\\
&\quad + \left|E_{\mathsf{II},k}\right|\left|E_{\mathsf{V},nmijk} - E'_{\mathsf{V},nmijk}\right| + \left|E'_{\mathsf{V},nmijk}\right|\left|E_{\mathsf{II},k} - E'_{\mathsf{II},k}\right|)\\
&\le 2 \times \frac{1}{10}\frac{C_{\phi'}M_\phi}{d}\left(\frac{1}{d}C_{\phi'}\|x - x'\|_{\max} + C_\phi\|\tilde{s} - \tilde{s}'\|_{\max}\right)\\
&\quad + \frac{C_{\phi'}M_\phi}{d}\frac{C_{\phi'}M_\phi}{d}\frac{C_\phi}{8}\left(\|x - x'\|_{\max} + \|s - s'\|_{\max}\right) + \frac{1}{4}\left(\frac{C_{\phi''}}{d}|x^j - (x')^j| + C_\phi\|\tilde{\tilde{s}} - \tilde{\tilde{s}}'\|_{\max}\right)\\
&\quad + \left(\frac{1}{d}C_{\phi''} + \frac{C_\phi}{1 - (C_\phi/4)}a_{AA,\max}\right)\frac{C_\phi}{10}(\|x - x'\|_{\max} + \|s - s'\|_{\max})\\
&\le \left(\frac{M_\phi C_{\phi'}^2}{5d^2} + \frac{C_\phi C_{\phi'}^2 M_\phi^2}{8d^2} + \frac{C_{\phi''}}{4d} + \frac{C_\phi}{10}\left(\frac{C_{\phi''}}{d} + \frac{C_\phi}{1 - (C_\phi/4)}a_{AA,\max}\right)\right)\|x - x'\|_{\max}\\
&\quad + \left(\frac{C_\phi C_{\phi'}^2 M_\phi^2}{8d^2} + \frac{C_\phi}{10}\left(\frac{C_{\phi''}}{d} + \frac{C_\phi}{1 - (C_\phi/4)}a_{AA,\max}\right)\right)\|s - s'\|_{\max}\\
&\quad + \frac{C_\phi M_\phi C_{\phi'}}{5d}\|\tilde{s} - \tilde{s}'\|_{\max} + \frac{C_\phi}{4}\|\tilde{\tilde{s}} - \tilde{\tilde{s}}'\|_{\max}.
\end{aligned}
\tag{3.38}
$$

$$
\begin{aligned}
&\left|F_{W^{nm},A^{ij},k}(u,v) - F_{W^{nm},A^{ij},k}(u',v')\right| \\
&\le \left|E_{\mathsf{I},k}E_{\mathsf{IV},nmk}\right|\left|E_{\mathsf{III},ijk} - E'_{\mathsf{III},ijk}\right| + \left|E_{\mathsf{I},k}E'_{\mathsf{III},ijk}\right|\left|E_{\mathsf{IV},nmk} - E'_{\mathsf{IV},nmk}\right| + \left|E'_{\mathsf{IV},nmk}E'_{\mathsf{III},ijk}\right|\left|E_{\mathsf{I},k} - E'_{\mathsf{I},k}\right| \\
&\quad + \left|E_{\mathsf{II},k}\right|\left|E_{\mathsf{VI},nmijk} - E'_{\mathsf{VI},nmijk}\right| + \left|E'_{\mathsf{VI},nmijk}\right|\left|E_{\mathsf{II},k} - E'_{\mathsf{II},k}\right|) \\
&\le \frac{1}{10}\times\frac{C_{\phi'}M_\phi}{N}\times\left(\frac{1}{d}C_{\phi'}\|x-x'\|_{\max} + C_\phi\|\tilde{s}-\tilde{s}'\|_{\max}\right) + \frac{1}{10}\times\frac{C_{\phi'}M_\phi}{d}\times\left(\frac{1}{N}C_{\phi'}\|x-x'\|_{\max} + C_\phi\|\tilde{s}-\tilde{s}'\|_{\max}\right) \\
&\quad + \frac{C_{\phi'}M_\phi}{N}\times\frac{C_{\phi'}M_\phi}{d}\times\frac{C_\phi}{8}\left(\|x-x'\|_{\max} + \|s-s'\|_{\max}\right) + \frac{1}{4}\left(\frac{1}{N}C_{\phi'}\|\tilde{s}-\tilde{s}'\|_{\max} + C_\phi\|\tilde{\tilde{s}}-\tilde{\tilde{s}}'\|_{\max}\right) \\
&\quad + \left(\frac{1}{Nd}\frac{C_{\phi'}^2}{4-C_\phi} + \frac{C_\phi}{1-(C_\phi/4)}a_{WA,\max}\right)\frac{C_\phi}{10}\left(\|x-x'\|_{\max} + \|s-s'\|_{\max}\right) \\
&\le \left(\frac{M_\phi C_{\phi'}^2}{5Nd} + \frac{C_\phi M_\phi^2 C_{\phi'}^2}{8Nd} + \frac{C_\phi}{10}\left(\frac{C_{\phi'}^2}{Nd(4-C_\phi)} + \frac{C_\phi}{1-(C_\phi/4)}a_{WA,\max}\right)\right)\|x-x'\|_{\max} \\
&\quad + \left(\frac{C_\phi C_{\phi'}^2 M_{\phi'}^2}{8Nd} + \frac{C_\phi}{10}\left(\frac{C_{\phi'}^2}{Nd(4-C_\phi)} + \frac{C_\phi}{1-(C_\phi/4)}a_{WA,\max}\right)\right)\|s-s'\|_{\max} \\
&\quad + \left(\frac{C_\phi M_\phi C_{\phi'}}{10N} + \frac{C_\phi M_\phi C_{\phi'}}{10d} + \frac{C_{\phi'}}{4N}\right)\|\tilde{s}-\tilde{s}'\|_{\max} + \frac{C_\phi}{4}\|\tilde{\tilde{s}}-\tilde{\tilde{s}}'\|_{\max}. \qquad (3.39)
\end{aligned}
$$

$$
\begin{aligned}
&\left|F_{W^{nm},W^{ij},k}(u,v) - F_{W^{nm},W^{ij},k}(u',v')\right| \\
&\le \left|E_{\mathsf{I},k}E_{\mathsf{IV},nmk}\right|\left|E_{\mathsf{IV},ijk} - E'_{\mathsf{IV},ijk}\right| + \left|E_{\mathsf{I},k}E'_{\mathsf{IV},ijk}\right|\left|E_{\mathsf{IV},nmk} - E'_{\mathsf{IV},nmk}\right| + \left|E'_{\mathsf{IV},nmk}E'_{\mathsf{IV},ijk}\right|\left|E_{\mathsf{I},k} - E'_{\mathsf{I},k}\right| \\
&\quad + \left|E_{\mathsf{II},k}\right|\left|E_{\mathsf{VII},nmijk} - E'_{\mathsf{VII},nmijk}\right| + \left|E'_{\mathsf{VII},nmijk}\right|\left|E_{\mathsf{II},k} - E'_{\mathsf{II},k}\right|) \\
&\le 2\times\frac{1}{10}\frac{C_{\phi'}M_\phi}{N}\left(\frac{1}{N}C_{\phi'}\|x-x'\|_{\max} + C_\phi\|\tilde{s}-\tilde{s}'\|_{\max}\right) + \frac{C_{\phi'}M_\phi}{N}\frac{C_{\phi'}M_\phi}{N}\frac{C_\phi}{8}\left(\|x-x'\|_{\max} + \|s-s'\|_{\max}\right) \\
&\quad + \frac{1}{4}\left(\frac{C_{\phi''}}{N}\|s-s'\|_{\max} + C_\phi\|\tilde{\tilde{s}}-\tilde{\tilde{s}}'\|_{\max}\right) + \left(\frac{C_{\phi''}}{N} + \frac{C_\phi}{1-(C_\phi/4)}a_{WW,\max}\right)\frac{C_\phi}{10}\left(\|x-x'\|_{\max} + \|s-s'\|_{\max}\right) \\
&\le \left(\frac{M_\phi C_{\phi'}^2}{5N^2} + \frac{C_\phi M_\phi^2 C_{\phi'}^2}{8N^2} + \frac{C_\phi}{10}\left(\frac{C_{\phi''}}{N} + \frac{C_\phi}{1-(C_\phi/4)}a_{WW,\max}\right)\right)\|x-x'\|_{\max} \\
&\quad + \left(\frac{C_\phi C_{\phi'}^2 M_{\phi'}^2}{8N^2} + \frac{C_{\phi''}}{4N} + \frac{C_\phi}{10}\left(\frac{C_{\phi''}}{N} + \frac{C_\phi}{1-(C_\phi/4)}a_{WW,\max}\right)\right)\|s-s'\|_{\max} \\
&\quad + \frac{C_\phi M_\phi C_{\phi'}}{5N}\|\tilde{s}-\tilde{s}'\|_{\max} + \frac{C_\phi}{4}\|\tilde{\tilde{s}}-\tilde{\tilde{s}}'\|_{\max}. \qquad (3.40)
\end{aligned}
$$

Gathering these estimates together, we obtain

$$
\begin{aligned}
&\max_{ijnmk}\left[|F_{A^{nm},A^{ij},k}(u,v) - F_{A^{nm},A^{ij},k}(u',v')| \vee |F_{W^{nm},A^{ij},k}(u,v) - F_{W^{nm},A^{ij},k}(u',v')|\right. \\
&\qquad \left. \vee |F_{W^{nm},W^{ij},k}(u,v) - F_{W^{nm},W^{ij},k}(u',v')|\right] \\
&\quad \le L_1\max\left[\|x-x'\|_{\max}, \|s-s'\|_{\max}, \|\tilde{s}-\tilde{s}'\|_{\max}, \|\tilde{\tilde{s}}-\tilde{\tilde{s}}'\|_{\max}\right],
\end{aligned}
$$

where $L_1$ is defined in Assumption 2.10. This concludes the proof of the lemma. □

In summary, we have

$$
\|F(u,v) - F(u',v')\|_{\max} \le L\|(u,v) - (u',v')\|_{\max}, \quad L = L_0 \vee L_1. \qquad (3.41)
$$

We therefore conclude by Proposition 3.1 that the transition kernel satisfies the contraction condition

$$\frac{\mathsf{Wass}_2(P((u,v,\eta),\cdot),P((u',v',\eta'),\cdot))}{\|(u,v,\eta)-(u',v',\eta')\|_{\max}} \le q := \sqrt{L^2+L_\wp^2}. \tag{3.42}$$

With this, we can formally establish the existence of invariant measure. Recall the definition of the induced transition operator $P^\vee\rho$, where

$$P^\vee\rho(A) = \int P(\mathsf{h},A)\,\rho(d\mathsf{h}).$$

**Proposition 3.10** ($P^\vee$ is a contraction)**.** *The transition operator $P^\vee$ is a contraction in $\mathcal{P}_2(\mathcal{X})$: for all $\rho,\rho' \in \mathcal{P}_2(\mathcal{X})$*

$$\mathsf{Wass}_2(P^\vee\rho,P^\vee\rho') \le q\,\mathsf{Wass}_2(\rho,\rho'). \tag{3.43}$$

*Proof.* This is [19, Proposition 14.3]. □

Further define the $t$-transition operators $P_t$ iteratively with

$$P^{t+1}(h,A) = \int_{\mathcal{R}} P(\mathsf{h},A)\,P^t(h,d\mathsf{h}), \quad P^0(h,A) = \delta_h(A), \tag{3.44}$$

alongside with the transition operators $(P^t)^\vee$.

**Theorem 3.11** (Existence of Invariant Measure)**.** *Let $\mu_t^\theta := (P^t)^\vee\rho$ be the distribution of $H_t$ (that depends on the initial distribution $\mu_0^\theta = \rho$). If $q^2 := L_\wp^2 + L^2 < 1$, then there exists a stationary distribution $\mu^\theta$ such that $P^\vee\mu^\theta = \mu^\theta$, and constant $C' > 0$ such that*

$$\mathsf{Wass}_2(\mu_t^\theta,\mu^\theta) \le C' q^t \overset{t\to\infty}{\to} 0. \tag{3.45}$$

*Proof.* We first note that $\mathsf{Wass}_2(\mu_1^\theta,\mu_0^\theta)$ is bounded. This is because for any couplings $\gamma$ between $\mu_1^\theta$ and $\mu_0^\theta$,

$$\begin{aligned}\int_{\mathcal{R}\times\mathcal{R}} \|\mathsf{h}-\mathsf{h}'\|_{\max}^2\gamma(d\mathsf{h},d\mathsf{h}') &\le 2\int_{\mathcal{R}\times\mathcal{R}}\left[\|\mathsf{h}\|_{\max}^2+\|\mathsf{h}'\|_{\max}^2\right]\gamma(d\mathsf{h},d\mathsf{h}')\\ &\le 2\left[\int_{\mathcal{R}}\|\mathsf{h}\|_{\max}^2\,\mu_1^\theta(d\mathsf{h})+\int_{\mathcal{R}}\|\mathsf{h}'\|_{\max}^2\,\mu_0^\theta(d\mathsf{h}')\right].\end{aligned}$$

for some constant $C > 0$ as $\mathcal{R}$ is bounded. By iteratively applying Proposition 3.10 and Banach Fixed Point Theorem, there exists a unique probability measure $\mu^\theta$ such that

$$P^\vee\mu^\theta = \mu^\theta, \tag{3.46}$$

i.e., $\rho^*$ is an invariant measure of the joint Markov chain $(U_t,V_t)$). Furthermore, we have

$$\mathsf{Wass}_2(\mu_t^\theta,\mu^\theta) = \mathsf{Wass}_2((P^\vee)^t\rho,\mu^\theta) \le \frac{q^t}{1-q}\mathsf{Wass}_2(\mu_1^\theta,\mu_0^\theta) = \frac{2C}{1-q}q^t. \tag{3.47}$$

□

### 3.3 Evaluating the limit loss and its derivative

This chapter aims to compute the limit of

$$L_T(\theta) = \frac{1}{2T}\sum_{t=1}^{T}(Y_t - f_t(\theta))^2 = \frac{1}{2T}\sum_{t=1}^{T}(f(X_t, Z_t, \eta_t) - \langle\lambda(B), S_t\rangle - \lambda(c))^2$$

as well as the derivative of the limits.

The limit is understood in the following sense:

**Lemma 3.12.** *Under Assumptions 2.3 and 2.10, we have* $\mathbb{E}[L_T(\theta)] \overset{T\to\infty}{\to} L(\theta)$*, where*

$$L(\theta) = \int \frac{1}{2}(f(\mathsf{x},\mathsf{z},\eta) - \langle\lambda(B),\mathsf{s}\rangle - \lambda(c))^2\,\mu^\theta(d\mathsf{h}), \quad \mathsf{h} = (\mathsf{x},\mathsf{z},\mathsf{s},\tilde{\mathsf{s}},\tilde{\tilde{\mathsf{s}}},\eta). \tag{3.48}$$

*Proof.* We define

$$\ell^\theta(h) = \frac{1}{2}(f(x,z,\eta) - \langle\lambda(B), s\rangle - \lambda(c))^2. \tag{3.49}$$

Then

$$\mathbb{E}\left[\frac{1}{2}(Y_t - f_t(\theta))^2\right] = \mathbb{E}\left[\frac{1}{2}(f(X_t,Z_t,\eta_t) - (\langle\lambda(B),S_t\rangle + \lambda(c)))^2\right] = \int \ell^\theta(\mathsf{h})\mu_t^\theta(d\mathsf{h}).$$

It therefore suffices to show that $\ell^\theta(h)$ is Lipschitz in $h$. Note that

$$\begin{aligned}
&|\ell^\theta(h) - \ell^\theta(h')|^2\\
&\le \frac{1}{2}\left|(f(x,z,\eta) - \langle\lambda(B),s\rangle - \lambda(c))^2 - (f(x',z',\eta') - \left\langle\lambda(B),s'\right\rangle - \lambda(c))^2\right|^2\\
&\le 6(L_f^2 + C_\lambda^2(N^2+1))[|f(x,z,\eta) - f(x',z',\eta')|^2 + N^2\|\lambda(B)\|_{\max}^2\|s-s'\|_{\max}^2]\\
&\le 6(L_f^2 + C_\lambda^2(N^2+1))^2\|h-h'\|_{\max}^2.
\end{aligned} \tag{3.50}$$

By Theorem 3.11, there exists a coupling $\gamma_t^\theta$ between $\mu_t^\theta$ and $\mu^\theta$ such that

$$\int \|\mathsf{h}-\mathsf{h}'\|_{\max}^2\,\gamma_t^\theta(d\mathsf{h},d\mathsf{h}') \le Cq^{2t}.$$

Therefore, one has

$$\begin{aligned}
\left|\int \ell^\theta(\mathsf{h})\mu_t^\theta(d\mathsf{h}) - \int \ell^\theta(\mathsf{h}')\mu^\theta(d\mathsf{h}')\right|^2 &\le \int |\ell^\theta(\mathsf{h}) - \ell^\theta(\mathsf{h}')|^2\,\gamma_t^\theta(d\mathsf{h},d\mathsf{h}')\\
&\le 6(L_f^2 + C_\lambda^2(N^2+1))^2\int \|\mathsf{h}-\mathsf{h}'\|_{\max}^2\,\gamma_t^\theta(d\mathsf{h},d\mathsf{h}')\\
&\le 6C(L_f^2 + C_\lambda^2(N^2+1))^2 q^{2t} \overset{t\to\infty}{\to} 0.
\end{aligned} \tag{3.51}$$

Finally, we have

$$\begin{aligned}
\left|\mathbb{E}[L_T(\theta)] - \int \ell^\theta(\mathsf{h}')\,\mu_t^\theta(d\mathsf{h}')\right| &\le \left|\frac{1}{T}\sum_{t=1}^{T}\int \ell^\theta(\mathsf{h})\,\mu_t^\theta(d\mathsf{h}) - \int \ell^\theta(\mathsf{h}')\,\mu^\theta(d\mathsf{h}')\right|\\
&\le \frac{1}{T}\sum_{t=1}^{T}\left|\int \ell^\theta(\mathsf{h})\,\mu_t^\theta(d\mathsf{h}) - \int \ell^\theta(\mathsf{h}')\,\mu^\theta(d\mathsf{h}')\right|\\
&\le \frac{1}{T}\sum_{t=1}^{T}\sqrt{6C(L_f^2 + C_\lambda^2(N^2+1))^2}q^t\\
&\le \frac{1}{T}\sqrt{6C(L_f^2 + C_\lambda^2(N^2+1))^2}\frac{q}{1-q} = O(1/T),
\end{aligned} \tag{3.52}$$

so $\mathbb{E}[L_T(\theta)] \overset{T\to\infty}{\to} L(\theta)$ uniformly in $\theta$. □

We can also use the compute the gradient and Hessian of $L(\theta)$. We start by recalling that for finite $T$,

$$\nabla_\theta L_T(\theta) = \nabla_\theta \left[ \frac{1}{2T} \sum_{t=1}^{T} (Y_t - f_t(\theta))^2 \right] = -\frac{1}{T} \sum_{t=1}^{T} (f_t(\theta) - Y_t) \nabla_\theta f_t(\theta), \tag{3.53}$$

where $\nabla_\theta f_t(\theta)$ is the gradient of $f$, flattened as vector and indexed by the components, e.g.

$$[\nabla_\theta f_t(\theta)]_{A^{ij}} = \frac{\partial f_t}{\partial A_{ij}} = \lambda(B)^\top \frac{\partial S_t}{\partial A_{ij}} = \lambda(B)^\top \tilde{S}_t^{A^{ij}}, \tag{3.54}$$

so on and so forth. Following similar computations, we see that

$$\nabla_\theta f_t(\theta) = \begin{bmatrix} \sum_{k=1}^N \lambda(B^k) \mathsf{vec}(\tilde{S}_t^{A,k}) \\ \sum_{k=1}^N \lambda(B^k) \mathsf{vec}(\tilde{S}_t^{W,k}) \\ \lambda'(B) \odot S_t \\ \lambda'(c) \end{bmatrix} \begin{matrix} \leftarrow \text{entries from } A \\ \leftarrow \text{entries from } W, \\ \leftarrow \text{entries from } B \end{matrix} \tag{3.55}$$

Furthermore, the Hessian is given as

$$\begin{aligned} \mathsf{Hess}_\theta L_T(\theta) &= \frac{1}{T} \sum_{t=1}^{T} \nabla_\theta \left[ -\big(Y_t - f_t(\theta)\big) \nabla_\theta f_t(\theta)^\top \right] \\ &= \frac{1}{T} \sum_{t=1}^{T} \left[ \nabla_\theta f_t(\theta) \nabla_\theta f_t(\theta)^\top - \big(Y_t - f_t(\theta)\big) \nabla_\theta^2 f_t(\theta)^\top \right], \end{aligned} \tag{3.56}$$

where

$$\mathsf{Hess}_\theta f_t(\theta) = \begin{bmatrix} \sum_{k=1}^N \lambda(B^k) \mathsf{flat}(\tilde{\tilde{S}}_t^{AA,k}) & \sum_{k=1}^N \lambda(B^k) \mathsf{flat}(\tilde{\tilde{S}}_t^{WA,k}) & \mathsf{flat}(\tilde{S}_t^A)^\top \mathsf{diag}(\lambda'(B)) & 0 \\ \sum_{k=1}^N \lambda(B^k) \mathsf{flat}(\tilde{\tilde{S}}_t^{AW,k}) & \sum_{k=1}^N \lambda(B^k) \mathsf{flat}(\tilde{\tilde{S}}_t^{WW,k}) & \mathsf{flat}(\tilde{S}_t^W)^\top \mathsf{diag}(\lambda'(B)) & 0 \\ \mathsf{diag}(\lambda'(B)) \mathsf{flat}(\tilde{S}_t^A) & \mathsf{diag}(\lambda'(B)) \mathsf{flat}(\tilde{S}_t^W) & \mathsf{diag}(\lambda''(B) \odot S_t) & 0 \\ 0 & 0 & 0 & \lambda''(c). \end{bmatrix} \tag{3.57}$$

We can verify this by checking the entries – for example, we have

$$\frac{\partial^2 f_t}{\partial B^k \partial A^{ij}} = \frac{\partial}{\partial B_k} \left[ \sum_{\ell=1}^N \lambda(B^\ell) \tilde{S}_t^{A^{ij},\ell} \right] = \lambda(B_k) \tilde{S}_t^{A^{ij},k}.$$

To compute $\nabla_\theta \mathbb{E}[L_T(\theta)]$ and $\mathsf{Hess}_\theta \mathbb{E}[L_T(\theta)]$, we need to exchange $\nabla_\theta$ and $\mathbb{E}$. This could be justified by noting that $\nabla_\theta L_T(\theta)$ and $\mathsf{Hess}_\theta L_T(\theta)$ are being bound (hence integrable). It follows from the following lemma:

**Lemma 3.13.** *Assume that Assumptions 2.3 and 2.10 hold. Then, there is a uniform constant $C < \infty$, independent of $t \in \mathbb{N}_+$, such that*

$$\|\nabla_\theta f_t(\theta)\|_{\max} \le C \tag{3.58}$$

$$\|\nabla_\theta f_t(\theta) \nabla_\theta f_t(\theta)^\top\|_{\max} \le C \tag{3.59}$$

$$\|\mathsf{Hess}_\theta f_t(\theta)\|_{\max} \le C, \tag{3.60}$$

*and consequently,*

$$\|\nabla_\theta f_t(\theta) \nabla_\theta f_t(\theta)^\top\|_{\mathsf{F}} \le C$$

$$\|\mathsf{Hess}_\theta f_t(\theta)\|_{\mathsf{F}} \le C,$$

*where $\|\cdot\|_{\mathsf{F}}$ is the Frobenious norm.*

*Proof.* The boundedness of $\|\nabla_\theta f_t(\nabla_\theta f_t)^\top\|_{\max}$ and $\|\mathsf{Hess}_\theta f_t\|_{\max}$ implies the boundedness of $\|\nabla_\theta f_t(\nabla_\theta f_t)^\top\|_{\mathsf{F}}$ and $\|\nabla^2_\theta f\|_{\mathsf{F}}$ respectively. This implies that it suffice to prove that each entries of $\nabla_\theta f_t$ is bounded. This follows from the boundedness of

- $\lambda(\cdot)$ and its derivatives by Assumption 2.10,
- $S_t$ since $\sigma(\cdot)$ is bounded,
- $\tilde{S}_t = (\tilde{S}^A_t, \tilde{S}^W_t)$ by Lemma 3.5, and
- $\tilde{\tilde{S}}_t = (\tilde{\tilde{S}}^{AA}_t, \tilde{\tilde{S}}^{WA}_t, \tilde{\tilde{S}}^{WW}_t)$ by Lemma 3.7.

□

**Lemma 3.14.** *Assume that Assumptions 2.3 and 2.10 hold. Then, we have that there is a uniform constant $C < \infty$, independent of $t \in \mathbb{R}_+$, such that*

$$\|\theta_{t+1} - \theta_t\|_{\max} \le C\alpha_t, \tag{3.61}$$

*and consequently*

$$\|\theta_{t+1} - \theta_t\|_2 \le C\alpha_t, \tag{3.62}$$

*for a potentially different finite constant $C < \infty$.*

*Proof.* This follows directly by the a-priori bounds of Lemma 3.5 and the online stochastic estimates (2.25). □

We can therefore exchange derivatives and the expectation operator $\mathbb{E}$ to compute $\nabla_\theta \mathbb{E}[L_T(\theta)] = \mathbb{E}[\nabla_\theta L_T(\theta)]$ and $\nabla^2_\theta \mathbb{E}[L_T(\theta)] = \mathbb{E}[\nabla^2_\theta L_T(\theta)]$. In particular, if we define the functions for $h = (x, z, s, \tilde{s}, \tilde{\tilde{s}}, \eta)$:

$$g^\theta(h) = g(h;\theta) := -(f(x,z,\eta) - \langle\lambda(B), s\rangle - \lambda(c)) \times \begin{bmatrix} \sum_k \lambda(B^k)\mathsf{vec}(\tilde{s}^{A,k}) \\ \sum_k \lambda(B^k)\mathsf{vec}(\tilde{s}^{W,k}) \\ \lambda'(B) \odot s \\ \lambda'(c) \end{bmatrix}, \tag{3.63}$$

then

$$\nabla_\theta \mathbb{E}[L_T(\theta)] = \mathbb{E}[\nabla_\theta L_T(\theta)] = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T -(Y_t - \langle\lambda(B), S_t\rangle - \lambda(c))\nabla_\theta f_t(\theta)\right] = \frac{1}{T}\sum_{t=1}^T \int g(\mathsf{h};\theta)\,\mu^\theta_t(d\mathsf{h}). \tag{3.64}$$

We further define the function

$$\begin{aligned} \mathrm{he}^\theta(h) = &\begin{bmatrix} \sum_k \lambda(B^k)\mathsf{vec}(\tilde{s}^{A,k}) \\ \sum_k \lambda(B^k)\mathsf{vec}(\tilde{s}^{W,k}) \\ \lambda'(B) \odot s \\ \lambda'(c) \end{bmatrix}\begin{bmatrix} \sum_k \lambda(B^k)\mathsf{vec}(\tilde{s}^{A,k}) \\ \sum_k \lambda(B^k)\mathsf{vec}(\tilde{s}^{W,k}) \\ \lambda'(B) \odot s \\ \lambda'(c) \end{bmatrix}^\top \\ &- (f(x,z,\eta) - \langle\lambda(B), s\rangle - \lambda(c)) \\ &\times \begin{bmatrix} \sum_k \lambda(B^k)\mathsf{flat}(\tilde{\tilde{s}}^{AA,k}) & \sum_k \lambda(B^k)\mathsf{flat}(\tilde{s}^{WA,k}) & \mathsf{flat}(\tilde{s}^A)\mathsf{diag}(\lambda'(B)) & 0 \\ \sum_k \lambda(B^k)\mathsf{flat}(\tilde{\tilde{s}}^{AW,k}) & \sum \lambda(B^k)\mathsf{flat}(\tilde{\tilde{s}}^{WW,k}) & \mathsf{flat}(\tilde{s}^W)^\top\lambda'(B) & 0 \\ \mathsf{diag}(\lambda'(B))\,\mathsf{flat}(\tilde{s}^A) & \mathsf{diag}(\lambda'(B)) \odot \mathsf{flat}(\tilde{s}^W) & \mathsf{diag}(\lambda''(B)) \odot s & 0 \\ 0 & 0 & 0 & \lambda''(c) \end{bmatrix}, \end{aligned}$$

then

$$\mathsf{Hess}_\theta \mathbb{E}[L_T(\theta)] = \mathbb{E}[\nabla_\theta^2 L_T(\theta)] = \int \mathrm{he}^\theta(\mathsf{h})\, \mu_t^\theta(d\mathsf{h}). \tag{3.65}$$

Due to the convergence rate $\mathsf{Wass}_1(\mu_t^\theta, \mu^\theta) \le Cq^t$ with $q < 1$, we can prove that this quantity has a limit as $T \to \infty$. In particular, we have Theorem 3.15.

**Lemma 3.15.** *Assume that Assumptions 2.3 and 2.10 hold. Then, we have that*

$$\lim_{T\to\infty} \nabla_\theta \mathbb{E}[L_T(\theta)] = \left\langle g^\theta(\mathsf{h}), \mu^\theta \right\rangle. \tag{3.66}$$

$$\lim_{T\to\infty} \mathsf{Hess}_\theta \mathbb{E}[L_T(\theta)] = \left\langle \mathrm{he}^\theta(\mathsf{h}), \mu^\theta \right\rangle. \tag{3.67}$$

*The limits are taken in max-norms.*

This result indicates that the online forward SGD algorithm (2.28) has a fluctuation term $\alpha_t(G_t - \nabla_\theta J(\theta_t))$ which should vanish as $t \to \infty$.

*Proof.* It suffices to prove that the functions $g^\theta(h)$ and $\mathrm{he}^\theta(h)$ are Lipschitz (assuming $h \in \mathcal{R}$). We note that

$$\begin{aligned}
\|g^\theta(h) - g^\theta(h')\|_{\max}^2 = \Bigg\| & (f(x,z,\eta) - \langle \lambda(B), s\rangle - \lambda(c)) \begin{bmatrix} \sum_k \lambda(B^k)\mathsf{vec}(\tilde{s}^{A,k}) \\ \sum_k \lambda(B^k)\mathsf{vec}(\tilde{s}^{W,k}) \\ \lambda'(B) \odot s \\ \lambda'(c) \end{bmatrix} \\
& - (f(x',z',\eta') - \langle \lambda(B), s'\rangle - \lambda(c) \begin{bmatrix} \sum_k \lambda(B^k)\mathsf{vec}[(\tilde{s}')^{A,k}] \\ \sum_k \lambda(B^k)\mathsf{vec}[(\tilde{s}')^{W,k}] \\ \lambda'(B) \odot s' \\ \lambda'(c) \end{bmatrix} \Bigg\|_{\max}^2 \\
\le\ & 2\left[L_f^2 |(x,z,\eta) - (x',z',\eta')|^2 + N^2 C_\lambda^2 \|s - s'\|_{\max}^2\right] \left[\max\left(\frac{C_\lambda C_{\phi'}}{\min(N,d)(4-C_\phi)}, C_{\lambda'}^2\right)\right]^2 \\
& + 3\left(L_f^2 + (N^2+1)C_\lambda^2\right) \left\| \begin{bmatrix} \sum_k \lambda(B^k)(\tilde{s}^{A,k} - (\tilde{s}')^{A,k}) \\ \sum_k \lambda(B^k)(\tilde{s}^{W,k} - (\tilde{s}')^{W,k}) \\ \lambda'(B) \odot (s - s') \\ 0 \end{bmatrix} \right\|_{\max}^2 \\
\le\ & C_{N,d}^{(1)} \|h - h'\|_{\max}^2,
\end{aligned} \tag{3.68}$$

where

$$\begin{aligned}
C_{N,d}^{(1)} = & \ 2\left(L_f^2 + C_\lambda^2 N^2\right) \left[\max\left(\frac{C_\lambda C_{\phi'}}{\min(N,d)(4-C_\phi)}, C_{\lambda'}^2\right)\right]^2 \\
& + 3\left(L_f^2 + (N^2+1)C_\lambda^2\right) \max(N^2, d^2) \max(C_\lambda, C_{\lambda'}).
\end{aligned} \tag{3.69}$$

Similarly, one could prove that there exists $C_{N,d}^{(2)} > 0$ such that

$$\|\mathrm{he}^\theta(h) - \mathrm{he}^\theta(h')\|_{\max}^2 \le C_{N,d}^{(2)} \|h - h'\|_{\max}^2. \tag{3.70}$$

Therefore, we could follow the arguments in the proof of Lemma 3.12 to establish the above limits. □

We note that the limits are uniform, so one could exchange $\nabla_\theta$ and $\lim_{T\to\infty}$ to establish that

**Theorem 3.16.**

$$\nabla_\theta L(\theta) = \nabla_\theta \left[ \lim_{T\to\infty} \mathbb{E}[L_T(\theta)] \right] = \lim_{T\to\infty} \nabla_\theta \mathbb{E}[L_T(\theta)] = \left\langle g^\theta(h), \mu^\theta \right\rangle. \tag{3.71}$$

$$\mathsf{Hess}_\theta L(\theta) = \mathsf{Hess}_\theta \left[ \lim_{T\to\infty} \mathbb{E}[L_T(\theta)] \right] = \lim_{T\to\infty} \nabla^2_\theta \mathbb{E}[L_T(\theta)] = \left\langle \mathrm{he}^\theta(h), \mu^\theta \right\rangle. \tag{3.72}$$

*Proof.* This follows from the fact that we have $\theta$-uniform convergence of $\nabla_\theta \mathbb{E}[L_T(\theta)]$ and $\mathsf{Hess}_\theta \mathbb{E}[L_T(\theta)]$ in $\theta$ (with rate $O(1/T)$), so we could exchange limit and derivatives, see e.g. Theorem 7.17 of [20]. □

As a corollary, we have

**Corollary 3.17.** *There exists constants $C > 0$ such that $\|\nabla_\theta L\|_{\max} \vee \|\mathsf{Hess}_\theta L\|_{\max} \le C$.*

*Proof.* This follows from the fact that $\|\nabla_\theta \mathbb{E}[L_T(\theta)]\|_{\max}$ and $\|\mathsf{Hess}_\theta \mathbb{E}[L_T(\theta)]\|_{\max}$ are both $T$-uniformly bounded. □

# 4 Poisson Equation

We will now construct a Poisson equation which will be used to bound the fluctuation term in (2.28) and prove it vanishes – at a suitably fast rate – as $t \to \infty$. We will emphasise the dependence of the process $H_t$ on the parameters $\theta$, thus adopting the notations $P_\theta$ for the transition kernel of $H^\theta$), and $P^t_\theta$ as the $t$-step transition probabilities of $H^\theta_t$. This will be crucial to the convergence analysis in the next section.

With the above notations, we can now define the Poisson equation.

**Definition 4.1** (Poisson equation)**.** We aim to find function $u(h;\theta)$ such that

$$u(h;\theta) - \int_{\mathcal{R}} u(\mathsf{h};\theta)\, P_\theta(h, d\mathsf{h}) = Q(h;\theta) - \int_{\mathcal{R}} Q(\mathsf{h};\theta)\, \mu^\theta(d\mathsf{h}), \tag{4.1}$$

where $h \in \mathcal{R}$ and $Q(h;\theta) : \mathcal{R} \to \mathbb{R}$ is a globally Lipschitz function (for each $\theta$).

**Lemma 4.2** (Existence of Solution)**.** *For any function $Q(h;\theta)$ which is globally Lipschitz in $h \in \mathcal{R}$ such that*

$$|Q(h;\theta) - Q(h';\theta)| \le L_Q \|h - h'\|_{\max} \tag{4.2}$$

*uniform in $\theta$, the following function $V(h;\theta)$ is a solution to the Poisson equation* (4.1) *for each $h \in \mathcal{R}$:*

$$V(h;\theta) = \sum_{t=0}^{\infty} \left( \int_{\mathcal{R}} Q(\mathsf{h};\theta)\, P^t_\theta(h, d\mathsf{h}) - \int_{\mathcal{R}} Q(\mathsf{h};\theta)\, \mu^\theta(d\mathsf{h}) \right). \tag{4.3}$$

*In addition, there is a finite constant $C < \infty$ that is uniform with respect to both $\theta \in \Theta$ and $h \in \mathcal{X}$ such that $|V(h;\theta)| \le C$.*

*Proof.* The proof follows that of Lemma 7.11 in [21]. We present the argument here emphasizing the differences. First, we prove that $V(h;\theta)$ is uniformly bounded. By Theorem 3.11 and [17, Theorem 4.1], there exists a constant $C > 0, q \in (0,1)$ (independent of $t$ and couplings $\gamma^\theta_t$ between $P^t_\theta(h,\cdot)$ and $\mu^\theta$ such that

$$\left[ \int \|\mathsf{h} - \mathsf{h}'\|^2_{\max}\, \gamma^\theta_t(d\mathsf{h}, d\mathsf{h}') \right]^{1/2} \le C q^t. \tag{4.4}$$

Consequently, by Lipschitzness of $Q$,

$$\begin{aligned}\left|\int_{\mathcal{R}} Q(\mathsf{h};\theta)\, P_\theta^t(h,d\mathsf{h}) - \int_{\mathcal{R}} Q(\mathsf{h}';\theta)\mu^\theta(d\mathsf{h}')\right| &= \left|\int_{\mathcal{R}\times\mathcal{R}} \big(Q(\mathsf{h};\theta) - Q(\mathsf{h}';\theta)\big)\,\gamma_t^\theta(d\mathsf{h},d\mathsf{h}')\right| \\ &\le \left[\int_{\mathcal{R}\times\mathcal{R}} \big|Q(\mathsf{h};\theta) - Q(\mathsf{h}';\theta)\big|^2\,\gamma_t^\theta(d\mathsf{h},d\mathsf{h}')\right]^{1/2} \\ &\le \left[\int_{\mathcal{R}\times\mathcal{R}} L_Q^2\,\big\|\mathsf{h}-\mathsf{h}'\big\|_{\max}^2\,\gamma_t^\theta(d\mathsf{h},d\mathsf{h}')\right]^{1/2} \\ &\le L_Q C q^t. \end{aligned} \tag{4.5}$$

Thus we have the following $\theta$-uniform bound,

$$|V(h;\theta)| \le \sum_{t=0}^{\infty}\left|\int_{\mathcal{R}} Q(\mathsf{h};\theta)\, P_\theta^t(h,\mathsf{h}) - \int_{\mathcal{R}} Q(\mathsf{h}';\theta)\,\mu^\theta(d\mathsf{h}')\right| \le \sum_{t=0}^{\infty} C L_Q q^t \le \frac{C L_Q}{1-q} < \infty. \tag{4.6}$$

We will now show that $V(h;\theta)$ is a solution to the Poisson equation (4.1). Starting with

$$\int_{\mathcal{R}} V(\mathsf{h};\theta)\, P_\theta^t(h,d\mathsf{h}) = \int_{\mathcal{R}}\left[\sum_{t=0}^{\infty}\left(\int_{\mathcal{R}} Q(\mathsf{h}';\theta)\, P_\theta^t(\mathsf{h},d\mathsf{h}') - \int_{\mathcal{R}} Q(\mathsf{h}';\theta)\,\mu^\theta(d\mathsf{h}')\right)\right] P_\theta(h,d\mathsf{h}'),$$

we can interchange the infinite sum and the integral due to Tonelli's theorem, which yields:

$$\begin{aligned}\int_{\mathcal{R}} V(\mathsf{h};\theta)\, P_\theta^t(h,d\mathsf{h}) &= \sum_{t=0}^{\infty}\left(\int_{\mathcal{R}}\int_{\mathcal{R}} Q(\mathsf{h}';\theta)\, P_\theta^t(\mathsf{h},d\mathsf{h}')\,\mathsf{P}_\theta(h,d\mathsf{h}) - \int_{\mathcal{R}}\int_{\mathcal{R}} Q(\mathsf{h}';\theta)\,\mu^\theta(d\mathsf{h}')\,\mathsf{P}_\theta(h,d\mathsf{h})\right) \\ &= \sum_{t=0}^{\infty}\left(\int_{\mathcal{R}} Q(\mathsf{h};\theta)\mathsf{P}_\theta^{t+1}(h,d\mathsf{h}) - \int_{\mathcal{R}} Q(\mathsf{h};\theta)\,\mu^\theta(d\mathsf{h})\right).\end{aligned}$$

Then,

$$\begin{aligned}V(h;\theta) - \int_{\mathcal{R}} V(\mathsf{h};\theta)\, P_\theta(h,d\mathsf{h}) &= \sum_{t=0}^{\infty}\left(\int_{\mathcal{R}} Q(\mathsf{h};\theta)\, P_\theta^t(h,d\mathsf{h}) - \int_{\mathcal{R}} Q(\mathsf{h};\theta)\,\mu^\theta(d\mathsf{h})\right) \\ &\quad - \sum_{t=0}^{\infty}\left(\int_{\mathcal{R}} Q(\mathsf{h};\theta)\, P_\theta^{t+1}(h,d\mathsf{h}) - \int_{\mathcal{R}} Q(\mathsf{h};\theta)\,\mu^\theta(d\mathsf{h})\right) \\ &= \int_{\mathcal{R}} Q(\mathsf{h};\theta)\, P_\theta^0(h,d\mathsf{h}) - \int_{\mathcal{R}} F(\mathsf{h};\theta)\,\mu^\theta(d\mathsf{h}) \\ &= Q(h;\theta) - \int_{\mathcal{R}} Q(\mathsf{h};\theta)\,\mu^\theta(d\mathsf{h}),\end{aligned} \tag{4.7}$$

which verifies that $V(h;\theta)$ is a solution to the Poisson equation (4.1). □

**Lemma 4.3** (Lipschitzness of solution)**.** *Let $Q(h;\theta)$ be a $\theta$-uniformly $L_Q$-Lipschitz function in $h$ as specified by equation* (4.2)*. Then, the solution $V(h;\theta)$ as defined in* (4.3) *for the Poisson equation* (4.1) *is $\theta$-uniformly Lipschitz globally in $h \in \mathcal{R}$.*

*Proof.* Recall that

$$V(h;\theta) = \sum_{t=0}^{\infty}\left(\int_{\mathcal{R}} Q(\mathsf{h};\theta)\, P_\theta^t(h,d\mathsf{h}) - \int_{\mathcal{R}} Q(\mathsf{h};\theta)\,\mu^\theta(d\mathsf{h})\right).$$

Therefore

$$\begin{aligned}\left|V(h;\theta)-V(h';\theta)\right| &= \left|\sum_{t=0}^{\infty}\left[\int_{\mathcal{R}} Q(\mathsf{h};\theta)\,P_\theta^t(h,d\mathsf{h}) - \int_{\mathcal{R}} Q(\mathsf{h}';\theta)\,P_\theta^t(h',d\mathsf{h}')\right]\right| \\ &\le \sum_{t=0}^{\infty}\left|\int_{\mathcal{R}} Q(\mathsf{h};\theta)\,P_\theta^t(h,d\mathsf{h}) - \int_{\mathcal{R}} Q(\mathsf{h}';\theta)\,P_\theta^t(h',d\mathsf{h}')\right|.\end{aligned}$$

By iteratively applying Proposition 3.10,

$$\mathsf{Wass}_2(P_\theta^t(h,\cdot),P_\theta^t(h',\cdot)) \le Cq^t\left|h-h'\right|_{\max} \tag{4.8}$$

for the same $C>0$ and $q\in(0,1)$ as specified in that Proposition. Thus by [17, Theorem 4.1], there are coupling $\gamma_{t,h,h'}^\theta$ between $P_\theta^t(h,\cdot)$ and $P_\theta^t(h',\cdot))$ such that

$$\left[\int_{\mathcal{R}\times\mathcal{R}} \|\mathsf{h}-\mathsf{h}'\|_{\max}^2\,\gamma_{t,h,h'}^\theta(d\mathsf{h},d\mathsf{h})\right]^{1/2} \le Cq^t\left\|h-h'\right\|_{\max},$$

Thus, as in the proof of Lemma 4.2,

$$\begin{aligned}\left|\int_{\mathcal{R}} Q(\mathsf{h};\theta)\,P_\theta^t(h,d\mathsf{h}) - \int_{\mathcal{R}} Q(\mathsf{h}';\theta)\,P_\theta^t(h',d\mathsf{h}')\right| &\le \left|\int_{\mathcal{R}\times\mathcal{R}}\left(Q(\mathsf{h};\theta)-Q(\mathsf{h}';\theta)\right)\gamma_{t,h,h'}^\theta(d\mathsf{h}',d\mathsf{h}')\right| \\ &\le \left[\int_{\mathcal{R}\times\mathcal{R}}\left(Q(\mathsf{h};\theta)-Q(\mathsf{h}';\theta)\right)^2\gamma_{t,h,h'}^\theta(d\mathsf{h}',d\mathsf{h}')\right]^{1/2} \\ &\le L_Q\left[\int_{\mathcal{R}\times\mathcal{R}}\|\mathsf{h}-\mathsf{h}'\|_{\max}^2\,\gamma_{t,h,h'}^\theta(d\mathsf{h},d\mathsf{h})\right]^{1/2} \\ &\le L_Q Cq^t\left\|h-h'\right\|_{\max}. \end{aligned}\tag{4.9}$$

Therefore, we have

$$\mathsf{Wass}_2(P_\theta^t(h,\cdot),P_\theta^t(h',\cdot)) \le \frac{L_Q C}{1-q}\left\|h-h'\right\|_{\max} \tag{4.10}$$

□

For our analysis, we would like to set

$$Q(h;\theta) = \nabla_\theta L(\theta)^\top g^\theta(h). \tag{4.11}$$

We shall first note that

**Lemma 4.4.** *The function $Q(h;\theta)$ is uniformly bounded in $h$ and $\theta$ for $h\in\mathcal{R}$.*

*Proof.* Since $\nabla_\theta L(\theta) = \langle g(h;\theta),\mu_\theta\rangle$ and $g(h;\theta)$ is uniformly bounded in $\theta$ for $h\in\mathcal{R}$, $\nabla_\theta L(\theta)$ is uniformly bounded in $\theta$. Consequently, for $h\in\mathcal{R}$, $\nabla_\theta L(\theta)^\top g(h;\theta)$ is uniformly bounded in $(h,\theta)\in\mathcal{R}\times\Theta$. □

By checking $Q$ is globally Lipschitz in $\mathcal{R}$, we can show that there exists a $\theta$-uniformly Lipschitz solution (global in $h$) for the Poisson equation (4.1).

**Lemma 4.5.** *There exists $L_Q>0$ such that the function $Q(h;\theta)$ is $\theta$-uniformly $L_Q$ globally Lipschitz in $h$ in both the $\|\cdot\|_{\max}$ and $\|\cdot\|_2$ norms.*

*Proof.* We present only the proof of the global Lipschitz property in the $\|\cdot\|_{\max}$ norm as the proof in the $\|\cdot\|_2$ follows the same way. We note that

$$\begin{aligned}|Q(h;\theta) - Q(h';\theta)| &= |\nabla_\theta L(\theta)^\top (g^\theta(h) - g^\theta(h'))| \\ &\le d_\theta \|\nabla_\theta L(\theta)\|_{\max} \|g^\theta(h) - g^\theta(h')\|_{\max} \\ &\le L_Q \|h - h'\|_{\max},\end{aligned}$$

thanks to the fact that $\|\nabla_\theta L(\theta)\|$ is bounded as proved in Corollary 3.17 and $g^\theta$ is Lipschitz as established through the calculations in Lemma 3.3. □

As a result, the following function

$$V(h;\theta) = \sum_{t=0}^{\infty} \left( \int_{\mathcal{R}} Q(\mathsf{h};\theta)\, P_\theta^t(h, d\mathsf{h}) - \int_{\mathcal{R}} F(\mathsf{h};\theta)\, \mu^\theta(d\mathsf{h}) \right) \tag{4.12}$$

is a well-defined, $\theta$-uniformly Lipschitz-in-$h$ solution to the Poisson equation. In fact, we will show that the function $V$ is differentiable in $\theta$ with uniformly-bounded derivatives.

To establish this, we first compute the derivatives

$$\frac{\partial}{\partial \diamondsuit} \mathbb{E}[Q(H_t;\theta)] = \frac{\partial}{\partial \diamondsuit} \int_{\mathcal{R}} Q(h;\theta)\, \mathsf{P}_t^\theta(h, d\mathsf{h}).$$

We shall start by computing the gradients $\nabla_\theta Q(h;\theta)$:

$$\nabla_\theta Q(h;\theta) = (\mathsf{Hess} L) g^\theta(h) + (D_\theta g^\theta(h))^\top \nabla_\theta L, \tag{4.13}$$

where $D_\theta g^\theta(h)$ is the (flatten) Jacobian matrix of $g^\theta(h)$ with respect to $\theta$. Computing the entries

$$\frac{\partial g^\theta}{\partial B_k} = -\lambda'(B^k) s^k \begin{bmatrix} \sum_\ell \lambda(B^\ell)\mathsf{vec}(\tilde{s}^{A,\ell}) \\ \sum_\ell \lambda(B^\ell)\mathsf{vec}(\tilde{s}^{W,\ell}) \\ \lambda'(B) \odot s \\ \lambda'(c) \end{bmatrix} + (f(x,z,\eta) - \langle \lambda(B), s\rangle - \lambda(c)) \begin{bmatrix} \lambda(B^k)\mathsf{vec}(\tilde{s}^{A,k}) \\ \lambda(B^k)\mathsf{vec}(\tilde{s}^{W,k}) \\ 0 \\ \lambda''(B^k) s^k \\ 0 \\ 0 \end{bmatrix}$$

$$\frac{\partial g^\theta}{\partial c} = -\lambda'(c) \begin{bmatrix} \sum_\ell \lambda(B^\ell)\mathsf{vec}(\tilde{s}^{A,\ell}) \\ \sum_\ell \lambda(B^\ell)\mathsf{vec}(\tilde{s}^{W,\ell}) \\ \lambda'(B) \odot s \\ \lambda'(c) \end{bmatrix} + (f(x,z,\eta) - \langle \lambda(B), s\rangle - \lambda(c)) \begin{bmatrix} 0 \\ 0 \\ 0 \\ \lambda''(c) \end{bmatrix}.$$

We have the following computation

$$\begin{aligned} D_\theta g^\theta(h) = (f(x,z,\eta) - \langle \lambda(B), x\rangle - \lambda(c)) &\begin{bmatrix} 0 & 0 & \mathsf{diag}(\lambda'(B))\mathsf{flat}(\tilde{s}^A) & 0 \\ 0 & 0 & \mathsf{diag}(\lambda'(B))\mathsf{flat}(\tilde{s}^W) & 0 \\ 0 & 0 & \mathsf{diag}(\lambda''(B) \odot s) & 0 \\ 0 & 0 & 0 & \lambda''(c) \end{bmatrix} \\ - &\begin{bmatrix} 0 & & & \\ & 0 & & \\ & & \mathsf{diag}(\lambda'(B) \odot s) & \\ & & & \lambda'(c) \end{bmatrix} \begin{bmatrix} \sum_\ell \lambda(B^\ell)\mathsf{vec}(\tilde{s}^{A,\ell}) \\ \sum_\ell \lambda(B^\ell)\mathsf{vec}(\tilde{s}^{W,\ell}) \\ \lambda'(B) \odot s \\ \lambda'(c) \end{bmatrix} \end{aligned}$$

so

$$\|\nabla_\theta Q(h;\theta)\| \le C_{N,d}\|\mathsf{Hess}\, L\|_{\max}\|g^\theta(h)\|_{\max} + C_{N,d}\|D_\theta g^\theta(h)\|_{\max}\|\nabla_\theta L\|_{\max} \le C,$$

so by dominated convergence theorem, we could exchange derivatives and expectations for the computation of derivatives.

We note that $P_\theta^t$ does not depend on the parameters $B$ and $C$, thus it is true that

$$\frac{\partial}{\partial \diamondsuit}\mathbb{E}[Q(H_t;\theta)] = \mathbb{E}\left[\frac{\partial Q(h;\theta)}{\partial \diamondsuit}\right], \quad \diamondsuit = B^j, C^j.$$

We also note that by exchanging derivatives and expectations, for variables $\diamondsuit = A^{ij}$ or $W^{ij}$

$$\begin{aligned}\frac{\partial}{\partial \diamondsuit}\mathbb{E}[Q(H_t;\theta)] &= \mathbb{E}\left[\frac{\partial Q}{\partial \diamondsuit}(H_t;\theta) + \sum_k \frac{\partial Q}{\partial s^k}(H_t;\theta)\tilde{S}_t^{\diamondsuit,k} + \sum_{\blacktriangle,k}\frac{\partial Q}{\partial \tilde{s}^{\blacktriangle,k}}(H_t;\theta)\tilde{\tilde{S}}_t^{\diamondsuit,\blacktriangle,k}\right] \\ &= \mathbb{E}\left[\frac{\partial Q}{\partial \diamondsuit} + \sum_k \frac{\partial Q}{\partial s^k}\tilde{S}_t^{\diamondsuit,k} + \sum_{\ell,\ell',k}\frac{\partial Q}{\partial \tilde{s}^{A^{\ell\ell'},k}}\tilde{\tilde{S}}_t^{\diamondsuit,A^{\ell\ell'},k} + \sum_{m,m',k}\frac{\partial Q}{\partial \tilde{s}^{W^{mm'},k}}\tilde{\tilde{S}}_t^{\diamondsuit,W^{mm'},k}\right] \quad (4.14)\end{aligned}$$

We shall define the vector $\mathfrak{i} := \mathfrak{i}(h;\theta)$ to be indexed by variables $\diamondsuit = A^{ij}, W^{ij}, B^i, C^i$ with the usual vectorisation convention in notation 2.13, such that

$$[\mathfrak{i}(h;\theta)]^\diamondsuit = \begin{cases}\sum_k \frac{\partial Q}{\partial s^k}\tilde{s}^{\diamondsuit,k} + \sum_{\ell,\ell',k}\frac{\partial Q}{\partial \tilde{s}^{A^{\ell\ell'},k}}\tilde{\tilde{s}}^{\diamondsuit,A^{\ell\ell'},k} + \sum_{m,m',k}\frac{\partial Q}{\partial \tilde{s}^{W^{mm'},k}}\tilde{\tilde{s}}^{\diamondsuit,W^{mm'},k} & \diamondsuit = A^{ij}, W^{ij} \\ 0 & \diamondsuit = B^j, C^j.\end{cases}$$

Let $\mathfrak{I} = \nabla_\theta Q + \mathfrak{i}$, then one could show by dominated convergence theorem,

$$\nabla_\theta \mathbb{E}[Q(H_t;\theta)] = \mathbb{E}[\mathfrak{I}(H_t;\theta)]. \quad (4.15)$$

**Lemma 4.6.** *Recall $Q(h;\theta) = \nabla_\theta L(\theta)^\top g(h;\theta)$, with*

$$g(h;\theta) = (f(x,z,\eta) - \langle\lambda(B), s\rangle - \lambda(c)) \times \begin{bmatrix}\sum_k \lambda(B^k)\tilde{s}^{A,k} \\ \sum_k \lambda(B^k)\tilde{s}^{W,k} \\ \lambda'(B)\odot s \\ \lambda'(c)\end{bmatrix},$$

*then $\mathfrak{I}(h;\theta)$ is globally Lipschitz in $h \in \mathcal{R}$ (with Lipschitz constant uniform in $\theta$).*

*Proof.* The Lipschitzness of $\nabla_\theta Q(h;\theta)$ follows from the boundedness of $\nabla_\theta L$, $\mathsf{Hess}\, L$, and the functions $g(\cdot;\theta)$ and $D_\theta g^\theta$ defined on a compact set $\mathcal{R}$ is smooth.

For the second term, we shall note that for $\heartsuit = s^k, \tilde{s}^{A^{ij},k}, \tilde{s}^{W^{ij},k}$,

$$\begin{aligned}\frac{\partial Q}{\partial \heartsuit} = \sum_\diamondsuit \frac{\partial L}{\partial \diamondsuit}\frac{[g]^\diamondsuit}{\partial \heartsuit} &= \sum_{\ell,\ell',l}\frac{\partial L}{\partial A^{\ell\ell'}}\lambda(B^l)\frac{\partial\left[(f(x,z,\eta) - \langle\lambda(B),s\rangle - \lambda(c))\,\tilde{s}^{A^{\ell\ell'},l}\right]}{\partial\heartsuit} \\ &+ \sum_{m,m',l}\frac{\partial L}{\partial W^{mm'}}\lambda(B^l)\frac{\partial\left[(f(x,z,\eta) - \langle\lambda(B),s\rangle - \lambda(c))\,\tilde{s}^{W^{mm'},l}\right]}{\partial\heartsuit} \\ &+ \sum_l \frac{\partial L}{\partial B^l}\lambda'(B^l)\frac{\partial\left[(f(x,z,\eta) - \langle\lambda(B),s\rangle - \lambda(c))\,s^l\right]}{\partial\heartsuit} \\ &+ \frac{\partial L}{\partial c}\lambda'(c)\frac{\partial\left[(f(x,z,\eta) - \langle\lambda(B),s\rangle - \lambda(c))\right]}{\partial\heartsuit}. \quad (4.16)\end{aligned}$$

In particular, we shall have

$$\frac{\partial Q}{\partial \tilde{s}^{A^{ij},k}} = \frac{\partial L}{\partial A^{ij}} \left(f(x,z,\eta) - \langle \lambda(B), s\rangle - \lambda(c)\right) \lambda(B^k) \tag{4.17}$$

$$\frac{\partial Q}{\partial \tilde{s}^{W^{ij},k}} = \frac{\partial L}{\partial W^{ij}} \left(f(x,z,\eta) - \langle \lambda(B), s\rangle - \lambda(c)\right) \lambda(B^k) \tag{4.18}$$

$$\begin{aligned}\frac{\partial Q}{\partial s^k} = -\lambda(B_k)\Bigg[ \sum_{\ell,\ell',l} \frac{\partial L}{\partial A^{\ell\ell'}} \lambda(B^l) \tilde{s}^{A^{\ell\ell'},l} + \sum_{m,m',l} \frac{\partial L}{\partial W^{mm'}} \lambda(B^l) \tilde{s}^{W^{mm'},l} + \sum_l \frac{\partial L}{\partial B^l} \lambda'(B^l) s^l + \frac{\partial L}{\partial c} \lambda'(c) \Bigg] \\ + \left(f(x,z,\eta) - \langle \lambda(B), s\rangle - \lambda(c)\right) \frac{\partial L}{\partial B^k} \lambda'(B^k).\end{aligned} \tag{4.19}$$

Since the entries are all smooth in $h$, and that $h \in \mathcal{R}$ (a compact set), so $\mathfrak{I}(h;\theta)$ is $h$-globally Lipschitz in $\mathcal{R}$. □

**Lemma 4.7.** *The solution $u(h;\theta)$ to the Poisson equation (4.1) is differentiable in $\theta$ and $\frac{\partial u}{\partial \theta}(h;\theta)$ is uniformly bounded with respect to $\theta \in \Theta$ and $h \in \mathcal{X}$.*

*Proof.* Recall that by Lemma 4.2 the following is a solution to the Poisson equation (4.1)

$$V(h;\theta) = \sum_{t=0}^{\infty} \left( \int_{\mathcal{R}} Q(\mathsf{h};\theta) P_\theta^t(h, d\mathsf{h}) - \int_{\mathcal{R}} Q(\mathsf{h};\theta) \mu^\theta(d\mathsf{h}) \right) = \sum_{t=0}^{\infty} \left[ \mathbb{E}[Q(H_t,\theta)] - \int_{\mathcal{R}} Q(\mathsf{h};\theta) \mu^\theta(d\mathsf{h}) \right].$$

We are interested in bounding $\nabla_\theta V(h;\theta)$. Define

$$\phi_t(h,\theta) = \mathbb{E}[Q(H_t,\theta)] - \int_{\mathcal{R}} Q(\mathsf{h};\theta) \mu^\theta(d\mathsf{h}).$$

Then,

$$\nabla_\theta \phi_t = \nabla_\theta \mathbb{E}[Q(H_t;\theta)] - \nabla_\theta \int_{\mathcal{R}} Q(\mathsf{h};\theta) \mu^\theta(d\mathsf{h}). \tag{4.20}$$

We have shown that $\nabla_\theta \mathbb{E}[Q(H_t;\theta)] = \mathbb{E}[\mathfrak{I}(H_t;\theta)]$. Since $\mathfrak{I}$ is globally $h$-Lipschitz with Lipschitz constant independent of $\theta$, we can make use of the uniform geometric convergence rate by Theorem 3.11 to exchange differentiation and integral signs. In fact, there shall be a coupling $\gamma_{h,t}^\theta$ between $P_\theta^t(h,\cdot)$ and $\mu^\theta(\cdot)$ such that

$$\left[\mathsf{Wass}_2(P_\theta^t(h,\cdot), \mu^\theta(\cdot))\right]^2 = \int \|\mathsf{h} - \mathsf{h}'\|_{\max}^2 \gamma_{h,t}^\theta(d\mathsf{h}, d\mathsf{h}') \le Cq^{2t}, \tag{4.21}$$

so

$$\begin{aligned}\left| \nabla_\theta \mathbb{E}[Q(H_t;\theta)] - \int_{\mathcal{R}} \mathfrak{I}(\mathsf{h};\theta)\, \mu^\theta(d\mathsf{h}) \right| &= \left| \int_{\mathcal{R}} \mathfrak{I}(\mathsf{h};\theta)\, P_\theta^t(h, d\mathsf{h}) - \int_{\mathcal{R}} \mathfrak{I}(\mathsf{h};\theta)\, \mu^\theta(d\mathsf{h}) \right| \\ &\le \sqrt{d_\theta} \left\| \int_{\mathcal{R}} \mathfrak{I}(\mathsf{h};\theta)\, P_\theta^t(h, d\mathsf{h}) - \int_{\mathcal{R}} \mathfrak{I}(\mathsf{h};\theta)\, \mu^\theta(d\mathsf{h}) \right\|_{\max} \\ &\le \sqrt{d_\theta} \int_{\mathcal{R}} \left\| \mathfrak{I}(\mathsf{h};\theta) - \mathfrak{I}(\mathsf{h}';\theta) \right\|_{\max} \gamma_{h,\theta}^\theta(d\mathsf{h}, d\mathsf{h}') \\ &\le C_{N,d}\, q^t \overset{t\to\infty}{\to} 0,\end{aligned} \tag{4.22}$$

where $C_{N,d}$ is the Lipschitz constant of the function $\mathfrak{I}$ as obtained in Lemma 4.6. The convergence is uniform, so:

$$\nabla_\theta \left[ \int_{\mathcal{R}} Q(\mathsf{h};\theta)\, \mu^\theta(d\mathsf{h}) \right] = \lim_{t\to\infty} \mathbb{E}\left[\mathfrak{I}(H_t;\theta)\right]. \tag{4.23}$$

The bound also shows that $\left|\frac{\partial \phi_t}{\partial \theta}(h,\theta)\right|$ is uniformly bounded with respect to $\theta \in \Theta$ and $h \in \mathcal{X}$. Let us now return to study $\frac{\partial V}{\partial \theta}(h;\theta)$.

$$\frac{\partial V}{\partial \theta}(h;\theta) = \frac{\partial}{\partial \theta}\left[\sum_{t=0}^{\infty} \phi_t(h;\theta)\right].$$

Due to the dominated convergence theorem,

$$\frac{\partial V}{\partial \theta}(h;\theta) = \sum_{t=0}^{\infty} \frac{\partial \phi_t}{\partial \theta}(h;\theta).$$

Due to the bound (4.22), we have the uniform bound

$$\left|\frac{\partial V}{\partial \theta}(h;\theta)\right| \leq C_{N,d},$$

where $C_{N,d}$ is independent of $\theta \in \Theta$ and $h \in \mathcal{R}$. □

# 5 A-priori Bounds for Convergence Analysis

Let us now consider the evolution of the loss function $L(\theta_t)$ during training:

$$\begin{aligned} L(\theta_{t+1}) - L(\theta_t) &= \nabla_\theta L(\theta_t)^\top (\theta_{t+1} - \theta_t) + (\theta_{t+1} - \theta_t)^\top \mathsf{Hess}_\theta L(\theta_t^*)(\theta_{t+1} - \theta_t) \\ &= -\alpha_t \nabla_\theta L(\theta_t)^\top \Big( \nabla_\theta L(\theta_t) - (G_t - \nabla_\theta L(\theta_t)) \Big) + R_t \\ &= -\alpha_t \nabla_\theta L(\theta_t)^\top \nabla_\theta L(\theta_t) + \alpha_t \nabla_\theta L(\theta_t)^\top (G_t - \nabla_\theta L(\theta_t)) + R_t, \end{aligned} \tag{5.1}$$

where $R_t = (\theta_{t+1} - \theta_t)^\top \mathsf{Hess}_\theta L(\theta_t^*)(\theta_{t+1} - \theta_t)$ and

$$G_t = -\Big( f(X_t, Z_t, \eta_t) - \big\langle \lambda(B_t), \bar{S}_t \big\rangle - \lambda(c_t) \Big) \times \begin{bmatrix} \sum_k \lambda(B_t^k) \hat{S}_t^{A,k} \\ \sum_k \lambda(B_t^k) \hat{S}_t^{W,k} \\ \lambda'(B_t) \odot \bar{S}_t \\ \lambda'(c_t) \end{bmatrix}.$$

We recall $Q(h;\theta) = \nabla_\theta L(\theta)^\top g(h;\theta)$ and

$$g(h;\theta) = \Big( f(x,z,\eta) - \langle \lambda(B), s \rangle - \lambda(c) \Big) \times \begin{bmatrix} \sum_k \lambda(B^k) \mathsf{vec}(\tilde{s}^{A,k}) \\ \sum_k \lambda(B^k) \mathsf{vec}(\tilde{s}^{W,k}) \\ \lambda'(B) \odot s \\ \lambda'(c) \end{bmatrix}.$$

If we define the joint process with online estimates of forward derivatives

$$\hat{H}_t = (X_t, Z_t, \bar{S}_t, \hat{S}_t, \eta_t) \tag{5.2}$$

Then, we can write

$$\begin{aligned} G_t &= g(\hat{H}_t; \theta_t), \\ \nabla_\theta L(\theta_t)^\top (G_t - \nabla_\theta L(\theta_t)) &= Q(\hat{H}_t, \theta_t) - \nabla_\theta L(\theta_t)^\top \nabla_\theta L(\theta_t). \end{aligned}$$

Due to equation (3.71), we can show that

$$\int Q(\mathsf{h},\theta)\,\mu_\theta(\mathsf{h}) = \int \left[\nabla_\theta L(\theta)^\top g(\mathsf{h};\theta)\right]\mu_\theta(d\mathsf{h}) = \nabla_\theta L(\theta)^\top\left[\int g(h;\theta)\,\mu_\theta(\mathsf{h})\right] = \nabla_\theta L(\theta)^\top \nabla_\theta L(\theta).$$

Therefore:

$$\alpha_t \nabla_\theta L(\theta_t)^\top (G_t - \nabla_\theta L(\theta_t)) = \alpha_t\left[Q(\hat{H}_t,\theta_t) - \int Q(\mathsf{h};\theta)\,\mu_{\theta_t}(d\mathsf{h})\right] \tag{5.3}$$

$$= \alpha_t\left[u(\hat{H}_t;\theta_t) - \int_{\mathcal{X}} u(\mathsf{h};\theta_t)\,P_{\theta_t}(\hat{H}_t,d\mathsf{h})\right], \tag{5.4}$$

where $u(h;\theta)$ is a solution to the Poisson equation (4.1), which is shown to be uniformly bounded.

We have shown in Lemmas 4.4 and 4.5 that the function $Q$ is bounded and Lipschitz in $h$. Using these bounds, we can now return to analyzing equation (5.1) using (5.4). In particular, summing (5.1) yields

$$L(\theta_T) = L(\theta_0) - \sum_{t=0}^{T-1}\alpha_t \nabla_\theta L(\theta_t)^\top \nabla_\theta L(\theta_t) + \sum_{t=0}^{T-1}\alpha_t\left[u(\hat{H}_t;\theta_t) - \int_{\mathcal{R}} u(\mathsf{h};\theta_t)\,P_{\theta_t}(\hat{H}_t,d\mathsf{h})\right] + \sum_{t=0}^{T-1} R_t.$$

We shall now prove the following proposition on the behavior of fluctuation and remainder terms.

**Proposition 5.1.** *Assume that Assumptions 2.3, 2.10 and 2.14 hold. There is a finite constant $C<\infty$ that is independent of $1\le T<\infty$ such that*

$$\sum_{t=0}^{T-1}\mathbb{E}\left|\alpha_t\left(u(\hat{H}_t;\theta_t) - \int_{\mathcal{R}} u(\mathsf{h};\theta_t)P_{\theta_t}(\hat{H}_t,\,d\mathsf{h})\right) + R_t\right| \le C.$$

The proof of Proposition 5.1 will be presented later in this section and it is a consequence of a number of intermediate results. In the lemmas that follow in this section, Assumptions 2.3, 2.10 and 2.14 will be assumed to hold throughout. The first step is to use the decomposition

$$\sum_{t=0}^{T-1}\alpha_t\left[u(\hat{H}_t;\theta_t) - \int_{\mathcal{R}} u(\mathsf{h};\theta_t)P_{\theta_t}(\hat{H}_t,d\mathsf{h})\right] = Q_T^1 + Q_T^2 + R_0^1, \tag{5.5}$$

where

$$\begin{aligned}
Q_T^1 &= \sum_{t=1}^{T-1}\alpha_t\left(u(\hat{H}_t;\theta_t) - \int_{\mathcal{R}} u(\mathsf{h};\theta_t)P_{\theta_t}(\hat{H}_{t-1},d\mathsf{h})\right),\\
Q_T^2 &= \sum_{t=1}^{T-1}\alpha_t\left(\int_{\mathcal{R}} u(\mathsf{h};\theta_t)P_{\theta_t}(\hat{H}_{t-1},d\mathsf{h}) - \int_{\mathcal{R}} u(\mathsf{h};\theta_t)P_{\theta_t}(\hat{H}_t,d\mathsf{h})\right),\\
R_j^1 &= \alpha_j\left(u(\hat{H}_j;\theta_j) - \int_{\mathcal{R}} u(\mathsf{h};\theta_j)\,P_{\theta_j}(\hat{H}_j,d\mathsf{h})\right).
\end{aligned} \tag{5.6}$$

Furthermore, we will decompose $Q_T^1$ into two more terms $Q_T^1 = Q_T^{1,A} + Q_T^{1,B}$, where

$$\begin{aligned}
Q_T^{1,A} &= \sum_{t=1}^{T-1}\alpha_t\left(\int_{\mathcal{R}} u(\mathsf{h};\theta_t)P_{\theta_{t-1}}(\hat{H}_{t-1},d\mathsf{h}) - \int_{\mathcal{R}} u(\mathsf{h};\theta_t)P_{\theta_t}(\hat{H}_{t-1},d\mathsf{h})\right),\\
Q_T^{1,B} &= \sum_{t=1}^{T-1}\alpha_t\left(u(\hat{H}_t;\theta_t) - \int_{\mathcal{R}} u(\mathsf{h};\theta_t)P_{\theta_{t-1}}(\hat{H}_{t-1},d\mathsf{h})\right),
\end{aligned} \tag{5.7}$$

Since $u(h;\theta)$ is uniformly bounded by Lemma 4.2, hence

$$|R_j^1| \le C\alpha_j < \infty, \tag{5.8}$$

where the $C$ does not depend upon $\theta$, $T$, or $j$.

In the lemmas that follow we analyze the terms $Q_T^{1,A}$, $Q_T^{1,B}$, $Q_T^2$ and $R_T$. These yield the proof of Proposition 5.1.

**Lemma 5.2.** *For $T > i$, define*

$$Q_{i,T}^{1,A} = \sum_{t=i}^{T-1} \alpha_t \left( \int_{\mathcal{R}} u(\mathsf{h};\theta_t) P_{\theta_{t-1}}(\hat{H}_{t-1}, d\mathsf{h}) - \int_{\mathcal{R}} u(\mathsf{h};\theta_t) P_{\theta_t}(\hat{H}_{t-1}, d\mathsf{h}) \right) = \sum_{t=i}^{T-1} Q^{1,A}(t). \tag{5.9}$$

*Note that $Q_T^{1,A} = Q_{1,T}^{1,A}$. Then, there are constants $C, C_1 < \infty$ independent of $T$ so that*

$$|Q_{i,T}^{1,A}| \le \sum_{t=i}^{T-1} |Q^{1,A}(t)| \le C < \infty,$$

*where $|Q^{1,A}(t)| \le C_1 \alpha_t^2$.*

*Proof.* From Lemma 4.3, $u(h;\theta)$ is globally Lipschitz in $h$ for $h \in \mathcal{X}$. Thus we have

$$\begin{aligned}
\left|Q^{1,A}(t)\right| = \Bigg| & \int u(\mathsf{x},\mathsf{z}, F_S^{\theta_{t-1}}(X_{t-1}, Z_{t-1}, \bar{S}_{t-1}, \hat{S}_{t-1}), F_{\tilde{S}}^{\theta_{t-1}}(X_{t-1}, Z_{t-1}, \bar{S}_{t-1}, \hat{S}_{t-1}); \theta_t) \wp(X_{t-1}, Z_{t-1}, d\mathsf{x}, d\mathsf{z}) \mu_\eta(d\eta) \\
& - \int u(\mathsf{x},\mathsf{z}, F_S^{\theta_t}(X_{t-1}, Z_{t-1}, \bar{S}_{t-1}, \hat{S}_{t-1}), F_{\tilde{S}}^{\theta_t}(X_{t-1}, Z_{t-1}, \bar{S}_{t-1}, \hat{S}_{t-1}); \theta_t) \wp(X_{t-1}, Z_{t-1}, d\mathsf{x}, d\mathsf{z}) \mu_\eta(d\eta) \Bigg| \\
\le C \Big( & \left\| F_S^{\theta_{t-1}}(X_{t-1}, Z_{t-1}, \bar{S}_{t-1}, \hat{S}_{t-1}) - F_S^{\theta_t}(X_{t-1}, Z_{t-1}, \bar{S}_{t-1}, \hat{S}_{t-1}) \right\|_{\max} \\
& \vee \left\| F_{\tilde{S}}^{\theta_{t-1}}(X_{t-1}, Z_{t-1}, \bar{S}_{t-1}, \hat{S}_{t-1}) - F_{\tilde{S}}^{\theta_t}(X_{t-1}, Z_{t-1}, \bar{S}_{t-1}, \hat{S}_{t-1}) \right\|_{\max} \Big)
\end{aligned}$$

We can now calculate bounds on the difference $\|F^{\theta_{t-1}}(\cdot) - F^{\theta_t}(\cdot)\|_{\max}$. In the calculations below, the finite constant $C$ may change from line to line.

$$\begin{aligned}
& \|F_S^{\theta_{t-1}}(X_{t-1}, Z_{t-1}, \bar{S}_{t-1}, \hat{S}_{t-1}) - F_S^{\theta_t}(X_{t-1}, Z_{t-1}, \bar{S}_{t-1}, \hat{S}_{t-1})\|_{\max} \\
& = \max_k \left| \sigma\left( \frac{1}{d} \left\langle \phi(A_{t-1}^{k,:}), X_{t-1} \right\rangle + \frac{1}{N} \left\langle \phi(W_{t-1}^{k,:}), \bar{S}_{t-1} \right\rangle \right) - \sigma\left( \frac{1}{d} \left\langle \phi(A_t^{k,:}), X_{t-1} \right\rangle + \frac{1}{N} \left\langle \phi(W_t^{k,:}), \bar{S}_{t-1} \right\rangle \right) \right| \\
& \le \max_k \frac{1}{4} \left[ \frac{1}{d} \sum_{\ell=1}^N \left| \phi(A_{t-1}^{k\ell}) - \phi(A_t^{k\ell}) \right| \left| X_{t-1}^\ell \right| + \frac{1}{N} \sum_{\ell=1}^N \left| \phi(W_{t-1}^{k\ell}) - \phi(W_t^{k\ell}) \right| \left| \bar{S}_{t-1}^\ell \right| \right] \\
& \le \max_k \frac{1}{4} \left[ \frac{1}{d} \sum_{\ell=1}^N C_{\phi'} \left| A_{t-1}^{k\ell} - A_t^{k\ell} \right| + \frac{1}{N} \sum_{\ell=1}^N C_{\phi'} \left| W_{t-1}^{k\ell} - W_t^{k\ell} \right| \right] \\
& \overset{(*)}{\le} C\alpha_{t-1},
\end{aligned} \tag{5.10}$$

where $(*)$ is from Lemma 3.14, which yields that $\|\theta_t - \theta_{t-1}\|_{\max} \le C\alpha_{t-1}$. Moreover,

$$
\begin{aligned}
&\|F^{\theta_{t-1}}_{A^{ij},k}(X_{t-1}, Z_{t-1}, \bar{S}_{t-1}, \hat{S}_{t-1}) - F^{\theta_t}_{A^{ij},k}(X_{t-1}, Z_{t-1}, \bar{S}_{t-1}, \hat{S}_{t-1})\|_{\max} \\
&= \left|\sigma'\left(\frac{1}{d}\left\langle\phi(A^{k,:}_{t-1}), X_{t-1}\right\rangle + \frac{1}{N}\left\langle\phi(W^{k,:}_{t-1}), \bar{S}_{t-1}\right\rangle\right) - \sigma'\left(\frac{1}{d}\left\langle\phi(A^{k,:}_{t}), X_{t-1}\right\rangle + \frac{1}{N}\left\langle\phi(W^{k,:}_{t}), \bar{S}_{t-1}\right\rangle\right)\right| \\
&\quad \times \left|\frac{1}{d}\delta_{ik}\phi'(A^{ij}_{t-1})X^j_{t-1} + \frac{1}{N}\sum_{\ell=1}^N \phi(W^{k,\ell}_{t-1})\hat{S}^{A^{ij},\ell}_{t-1}\right| \\
&\quad + \left|\sigma'\left(\frac{1}{d}\left\langle\phi(A^{k,:}_{t}), X_{t-1}\right\rangle + \frac{1}{N}\left\langle\phi(W^{k,:}_{t}), \bar{S}_{t-1}\right\rangle\right)\right|\left|\frac{1}{d}\delta_{ik}(\phi'(A^{ij}_{t-1}) - \phi'(A^{ij}_{t}))X^j_{t-1} + \frac{1}{N}\sum_{\ell=1}^N(\phi(W^{k\ell}_{t-1}) - \phi(W^{k\ell}_{t}))\hat{S}^{A^{ij},\ell}_{t-1}\right| \\
&\le \frac{3}{10}\left|\frac{1}{d}\sum_{\ell=1}^N(\phi(A^{k\ell}_{t-1}) - \phi(A^{k\ell}_{t}))X^\ell_{t-1} + \frac{1}{N}\sum_{\ell=1}^N(\phi(W^{k\ell}_{t-1}) - \phi(W^{k\ell}_{t}))\bar{S}^\ell_{t-1}\right| \\
&\quad \times \frac{1}{4}\left|\frac{1}{d}\delta_{ik}(\phi'(A^{ij}_{t-1}) - \phi'(A^{ij}_{t}))X^j_{t-1} + \frac{1}{N}\sum_{\ell=1}^N(\phi(W^{k\ell}_{t-1}) - \phi(W^{k\ell}_{t}))\hat{S}^{A^{ij},\ell}_{t-1}\right| \\
&\overset{(*)}{\le} C\alpha_{t-1}.
\end{aligned} \tag{5.11}
$$

$$
\begin{aligned}
&\|F^{\theta_{t-1}}_{W^{ij},k}(X_{t-1}, Z_{t-1}, \bar{S}_{t-1}, \hat{S}_{t-1}) - F^{\theta_t}_{A^{ij},k}(X_{t-1}, Z_{t-1}, \bar{S}_{t-1}, \hat{S}_{t-1})\|_{\max} \\
&= \left|\sigma'\left(\frac{1}{d}\left\langle\phi(A^{k,:}_{t-1}), X_{t-1}\right\rangle + \frac{1}{N}\left\langle\phi(W^{k,:}_{t-1}), \bar{S}_{t-1}\right\rangle\right) - \sigma'\left(\frac{1}{d}\left\langle\phi(A^{k,:}_{t}), X_{t-1}\right\rangle + \frac{1}{N}\left\langle\phi(W^{k,:}_{t}), \bar{S}_{t-1}\right\rangle\right)\right| \\
&\quad \times \left|\frac{1}{d}\delta_{ik}\phi'(W^{ij}_{t-1})\bar{S}^j_{t-1} + \frac{1}{N}\sum_{\ell=1}^N \phi(W^{k,\ell}_{t-1})\hat{S}^{W^{ij},\ell}_{t-1}\right| \\
&\quad + \left|\sigma'\left(\frac{1}{d}\left\langle\phi(A^{k,:}_{t}), X_{t-1}\right\rangle + \frac{1}{N}\left\langle\phi(W^{k,:}_{t}), \bar{S}_{t-1}\right\rangle\right)\right|\left|\frac{1}{d}\delta_{ik}(\phi'(W^{ij}_{t-1}) - \phi'(W^{ij}_{t}))X^j_{t-1} + \frac{1}{N}\sum_{\ell=1}^N(\phi(W^{k\ell}_{t-1}) - \phi(W^{k\ell}_{t}))\hat{S}^{W^{ij},\ell}_{t-1}\right| \\
&\le \frac{3}{10}\left|\frac{1}{d}\sum_{\ell=1}^N(\phi(A^{k\ell}_{t-1}) - \phi(A^{k\ell}_{t}))X^\ell_{t-1} + \frac{1}{N}\sum_{\ell=1}^N(\phi(W^{k\ell}_{t-1}) - \phi(W^{k\ell}_{t}))\bar{S}^\ell_{t-1}\right| \\
&\quad \times \frac{1}{4}\left|\frac{1}{d}\delta_{ik}(\phi'(W^{ij}_{t-1}) - \phi'(W^{ij}_{t}))\bar{S}^j_{t-1} + \frac{1}{N}\sum_{\ell=1}^N(\phi(W^{k\ell}_{t-1}) - \phi(W^{k\ell}_{t}))\hat{S}^{W^{ij},\ell}_{t-1}\right| \\
&\overset{(*)}{\le} C\alpha_{t-1}.
\end{aligned} \tag{5.12}
$$

Consequently, we have for some finite constant $C < \infty$ that may change from line to line

$$
\begin{aligned}
|Q^{1,A}_{i,T}| &\le \sum_{t=i}^{T-1} |Q^{1,A}(t)| \\
&\le C\sum_{t=i}^{T-1} \alpha_t\alpha_{t-1} \\
&\le C < \infty.
\end{aligned} \tag{5.13}
$$

□

**Lemma 5.3.** *Define*

$$
M_t = u(H_t;\theta_t) - \int_{\mathcal{X}} u(h';\theta_t)\, P_{\theta_{t-1}}(\hat{H}_{t-1}, d\mathsf{h}), \tag{5.14}
$$

*and for $T > j$,*

$$Q_{j,T}^{1,B} = \sum_{t=j}^{T-1} \alpha_t M_t. \tag{5.15}$$

*Note that $Q_T^{1,B} = Q_{1,T}^{1,B}$. Then, there are constants $C, C_1 < \infty$ independent of $T$ so that*

$$\mathbb{E}\left[(Q_{j,T}^{1,B})^2\right] \leq C_1 \sum_{t=j}^{T-1} \alpha_t^2 \leq C < \infty.$$

*The constant $C$ does not depend upon $T$ or $j$.*

*Proof.* We start by writing

$$\begin{aligned}
\mathbb{E}\left[(Q_{j,T}^{1,B})^2\right] &= \mathbb{E}\left[\sum_{i=j}^{T-1}\sum_{t=j}^{T-1} \alpha_t\alpha_i M_t M_i\right] \\
&= \sum_{i=j}^{T-1}\sum_{t=j}^{T-1} \alpha_t\alpha_i \mathbb{E}[M_t M_i] \\
&= \sum_{i=j}^{T-1} \sum_{t>i, t\leq T-1} \alpha_t\alpha_i \mathbb{E}[M_t M_i] + \sum_{t=j}^{T-1} \sum_{i>t, i\leq T-1} \alpha_t\alpha_i \mathbb{E}[M_t M_i] + \sum_{t=j}^{T-1} \alpha_t^2 \mathbb{E}[M_t^2].
\end{aligned} \tag{5.16}$$

Note that the joint process $(\hat{H}_t, \theta_t)$ is Markov (in fact $\theta_{t+1} = \theta_t - \alpha_t g(\hat{H}_t; \theta_t)$ is a deterministic function of $(\hat{H}_t, \theta_t)$). Using iterated expectations, we can show that if $t > i$:

$$\begin{aligned}
\mathbb{E}\left[M_t \middle| \hat{H}_{i-1}, \theta_{i-1}\right] &= \mathbb{E}\left[\mathbb{E}\left[M_t \middle| \hat{H}_{t-1}, \theta_{t-1}\right] \middle| \hat{H}_{i-1}, \theta_{i-1}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[u(\hat{H}_t; \theta_t) - \int_{\mathcal{R}} u(\mathsf{h}; \theta_t) P_{\theta_{t-1}}(H_{t-1}, d\mathsf{h}) \middle| \hat{H}_{t-1}, \theta_{t-1}\right] \middle| \hat{H}_{i-1}, \theta_{i-1}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[u(\hat{H}_t; \theta_t) \middle| \hat{H}_{t-1}, \theta_{t-1}\right] - \int_{\mathcal{R}} u(\mathsf{h}; \theta_t)\, P_{\theta_{t-1}}(\hat{H}_{t-1}, d\mathsf{h}) \middle| \hat{H}_{i-1}, \theta_{i-1}\right].
\end{aligned} \tag{5.17}$$

We note that

$$\begin{aligned}
&\mathbb{E}\left[u(\hat{H}_t; \theta_t) \middle| \hat{H}_{t-1}, \theta_{t-1}\right] \\
&= \mathbb{E}\left[u(X_t, Z_t, F_S^{\theta_{t-1}}(X_{t-1}, Z_{t-1}, \bar{S}_{t-1}, \hat{S}_{t-1}), F_{\tilde{S}}^{\theta_{t-1}}(X_{t-1}, Z_{t-1}, \bar{S}_{t-1}, \hat{S}_{t-1}); \theta_t) \middle| \hat{H}_{t-1}, \theta_{t-1}\right] \\
&= \int u(\mathsf{x}, \mathsf{z}, F_S^{\theta_{t-1}}(X_{t-1}, Z_{t-1}, \bar{S}_{t-1}, \hat{S}_{t-1}), F_{\tilde{S}}^{\theta_{t-1}}(X_{t-1}, Z_{t-1}, \bar{S}_{t-1}, \hat{S}_{t-1}); \theta_t))\, \wp(X_{t-1}, Z_{t-1}, d\mathsf{x}, d\mathsf{z}) \\
&= \int u(\mathsf{h}; \theta_t) P_{\theta_{t-1}}(\hat{H}_{t-1}, d\mathsf{h}).
\end{aligned} \tag{5.18}$$

Therefore for all $t > i$,

$$\mathbb{E}\left[M_t \middle| \hat{H}_{i-1}, \theta_{i-1}\right] = 0. \tag{5.19}$$

Using iterated expectations again,

$$\begin{aligned}
\sum_{i=j}^{T-1} \sum_{t>i, t\leq T-1} \alpha_t\alpha_i \mathbb{E}[M_t M_i] &= \sum_{i=j}^{T-1} \sum_{t>i, t\leq T-1} \alpha_t\alpha_i \mathbb{E}\left[\mathbb{E}\left[M_t M_i \middle| \hat{H}_{i-1}, \theta_{i-1}\right]\right] \\
&= \sum_{i=j}^{T-1} \sum_{t>i, t\leq T-1} \alpha_t\alpha_i \mathbb{E}\left[M_i \mathbb{E}\left[M_t \middle| \hat{H}_{i-1}, \theta_{i-1}\right]\right] \\
&= 0.
\end{aligned} \tag{5.20}$$

Similarly, we have

$$\sum_{t=1}^{T-1} \sum_{i>t, i\le T-1} \alpha_t \alpha_i \mathbb{E}[M_t M_i] = 0. \tag{5.21}$$

Since by Lemma 4.2 $u(h;\theta)$ is uniformly bounded, $\mathbb{E}[M_t^2] < C < \infty$, where $C$ is uniform in $t$. Therefore, since $\sum_{t=0}^{\infty} \alpha_t^2 < \infty$,

$$\mathbb{E}\Big[(Q_{j,T}^{1,B})^2\Big] \le \sum_{t=j}^{T-1} \alpha_t^2 \mathbb{E}[M_t^2] \le C \sum_{t=j}^{T-1} \alpha_t^2 < \infty, \tag{5.22}$$

which proves the uniform bound (5.16). □

**Lemma 5.4.** *Define*

$$Q_{j,T}^2 = \sum_{t=j}^{T-1} \alpha_t \left( \int_{\mathcal{R}} u(h';\theta_t)\, P_{\theta_t}(\hat{H}_{t-1}, d\mathsf{h}) - \int_{\mathcal{R}} u(\mathsf{h};\theta_t)\, P_{\theta_t}(\hat{H}_t, d\mathsf{h}) \right). \tag{5.23}$$

*Note that $Q_T^2 = Q_{1,T}^2$. Then, there are constants $C, C_1 < \infty$ independent of $T$ so that*

$$|Q_{j,T}^2| \le C_1 \sum_{t=j}^{T-2} \alpha_{t+1}\alpha_t + \alpha_j C_2 \le C < \infty. \tag{5.24}$$

*The constant $C$ does not depend upon $T$ or $j$.*

*Proof.* We begin by decomposing $Q_{j,T}^2$ into several terms:

$$\begin{aligned}
Q_{j,T}^2 &= \sum_{t=j}^{T-1} \alpha_t \left( \int_{\mathcal{R}} u(\mathsf{h};\theta_t) P_{\theta_t}(\hat{H}_{t-1}, d\mathsf{h}) - \int_{\mathcal{R}} u(\mathsf{h};\theta_t) P_{\theta_t}(\hat{H}_t, d\mathsf{h}) \right) \\
&= \alpha_j \int_{\mathcal{R}} u(\mathsf{h};\theta_j) P_{\theta_j}(\hat{H}_{j-1}, d\mathsf{h}) + \sum_{t=j+1}^{T-1} \alpha_t \int_{\mathcal{R}} u(\mathsf{h};\theta_t) P_{\theta_t}(\hat{H}_{t-1}, d\mathsf{h}) - \sum_{t=j}^{T-1} \alpha_t \int_{\mathcal{R}} u(\mathsf{h};\theta_t) P_{\theta_t}(\hat{H}_t, d\mathsf{h}) \\
&= \alpha_j \int_{\mathcal{R}} u(\mathsf{h};\theta_j) P_{\theta_j}(\hat{H}_{j-1}, d\mathsf{h}) + \sum_{t=j}^{T-2} \alpha_{t+1} \int_{\mathcal{R}} u(\mathsf{h};\theta_{t+1}) P_{\theta_{t+1}}(\hat{H}_t, d\mathsf{h}) - \sum_{t=j}^{T-1} \alpha_t \int_{\mathcal{R}} u(\mathsf{h};\theta_t) P_{\theta_t}(\hat{H}_t, d\mathsf{h}) \\
&= \sum_{t=j}^{T-2} \left[ \alpha_{t+1} \int_{\mathcal{R}} u(\mathsf{h};\theta_{t+1}) P_{\theta_{t+1}}(\hat{H}_t, d\mathsf{h}) - \alpha_t \int_{\mathcal{R}} u(\mathsf{h};\theta_t) P_{\theta_t}(\hat{H}_t, d\mathsf{h}) \right] + R_{j,T}^2 \\
&= \sum_{t=j}^{T-2} \alpha_{t+1} \int_{\mathcal{R}} u(\mathsf{h};\theta_{t+1}) \big( P_{\theta_{t+1}}(\hat{H}_t, d\mathsf{h}) - P_{\theta_t}(\hat{H}_t, d\mathsf{h}) \big) + \sum_{t=j}^{T-2} (\alpha_{t+1} - \alpha_t) \int_{\mathcal{R}} u(\mathsf{h};\theta_{t+1}) P_{\theta_t}(\hat{H}_t, d\mathsf{h}) \\
&\quad + \sum_{t=j}^{T-2} \alpha_t \int_{\mathcal{R}} \big( u(\mathsf{h};\theta_{t+1}) - u(\mathsf{h};\theta_t) \big) P_{\theta_t}(\hat{H}_t, d\mathsf{h}) + R_{j,T}^2,
\end{aligned} \tag{5.25}$$

where

$$R_{j,T}^2 = \alpha_j \int_{\mathcal{X}} u(h';\theta_j) P_{\theta_j}(dh'|H_{j-1}) + \alpha_{T-1} \int_{\mathcal{X}} u(\mathsf{h};\theta_{T-1}) P_{\theta_{T-1}}(\hat{H}_{T-1}, d\mathsf{h}). \tag{5.26}$$

Since by Lemma 4.2 $u(h;\theta)$ is uniformly bounded by some constant $C < +\infty$ and $\alpha_t$ is monotonically decreasing,

$$|R_{j,T}^2| \le C\alpha_j. \tag{5.27}$$

For the same reason, we shall also have

$$\left|\sum_{t=j}^{T-2}(\alpha_{t+1}-\alpha_t)\int_{\mathcal{R}}u(\mathsf{h};\theta_{t+1})P_{\theta_t}(\hat{H}_t,d\mathsf{h})\right|\le C\sum_{t=j}^{T-2}(\alpha_t-\alpha_{t+1})\le C\alpha_j.$$

Recall from equation (3.62) that $\|\theta_{t+1}-\theta_t\|_{\max}<C\alpha_t$. Due to the uniform bound from Lemma 4.7 and the Remark 2.15 $\sum_{t=1}^{\infty}\alpha_{t+1}\alpha_t<\infty$, we can prove the second term is bounded:

$$\begin{aligned}\left|\sum_{t=j}^{T-2}\alpha_{t+1}\int_{\mathcal{R}}\big(u(\mathsf{h};\theta_{t+1})-u(\mathsf{h};\theta_t)\big)P_{\theta_t}(H_t,d\mathsf{h})\right|&\le\sum_{t=j}^{T-2}\alpha_{t+1}\int_{\mathcal{R}}C\underbrace{\left\|\frac{\partial u}{\partial\theta}(\mathsf{h};\theta_t^*)\right\|_{\max}}_{\le C\text{ (by Lemma 4.7)}}\underbrace{\|\theta_{t+1}-\theta_t\|_{\max}}_{\le C\alpha_t}P_{\theta_t}(\hat{H}_t,d\mathsf{h})\\&\le C\sum_{t=j}^{T-2}\alpha_{t+1}\alpha_t\le C<\infty.\end{aligned}$$

The first term can also be bounded in the same way as Lemma 5.2, noticing that

$$\begin{aligned}&\left|\int_{\mathcal{R}}u(\mathsf{h};\theta_{t+1})\big(P_{\theta_{t+1}}(\hat{H}_t,d\mathsf{h})-P_{\theta_t}(\hat{H}_t,d\mathsf{h})\big)\right|\\&=\left|\int u(\mathsf{x},\mathsf{z},F_S^{\theta_{t+1}}(X_t,Z_t,\bar{S}_t,\hat{S}_t),F_{\tilde{S}}^{\theta_{t+1}}(X_t,Z_t,\bar{S}_t,\hat{S}_t);\theta_{t+1})\wp(X_t,Z_t,d\mathsf{x},d\mathsf{z})\mu_\eta(d\eta)\right.\\&\qquad\left.-\int u(\mathsf{x},\mathsf{z},F_S^{\theta_t}(X_t,Z_t,\bar{S}_t,\hat{S}_t),F_{\tilde{S}}^{\theta_t}(X_t,Z_t,\bar{S}_t,\hat{S}_t);\theta_{t+1})\wp(X_t,Z_t,d\mathsf{x},d\mathsf{z})\mu_\eta(d\eta)\right|\\&\le C\Big(\left\|F_S^{\theta_{t+1}}(X_t,Z_t,\bar{S}_t,\hat{S}_t)-F_S^{\theta_t}(X_t,Z_t,\bar{S}_t,\hat{S}_t)\right\|_{\max}\vee\left\|F_{\tilde{S}}^{\theta_{t+1}}(X_t,Z_t,\bar{S}_t,\hat{S}_t)-F_{\tilde{S}}^{\theta_t}(X_t,Z_t,\bar{S}_t,\hat{S}_t)\right\|_{\max}\Big),\end{aligned}$$

as $u(h;\theta)$ is globally Lipschitz in $h$ (with Lipschitz constant $L_u$ uniform in $\theta$) by Lemma 4.3. Consequently,

$$\left|\sum_{t=j}^{T-2}\alpha_{t+1}\int_{\mathcal{R}}u(\mathsf{h};\theta_{t+1})\big(P_{\theta_{t+1}}(\hat{H}_t,d\mathsf{h})-P_{\theta_t}(\hat{H}_t,d\mathsf{h})\big)\right|\le C\sum_{t=j}^{T-2}\alpha_{t+1}\alpha_t<\infty.\tag{5.28}$$

The finite constant $C<\infty$ may change from line to line. Combining these bounds proves the uniform bound (5.24). □

**Lemma 5.5.** *For any $j<T$, the remainder term $R_t$ satisfies*

$$\sum_{t=j}^{T-1}|R_t|\le C_1\sum_{t=j}^{T-1}\alpha_t^2\le C<\infty.$$

*The constant $C$ does not depend upon $T$ or $j$.*

*Proof.* We recall that $R_t=(\theta_{t+1}-\theta_t)^\top\mathsf{Hess}_\theta L(\theta_t^*)(\theta_{t+1}-\theta_t)$ where $H_\theta(\theta)$ is the Hessian matrix of $L(\theta)$. Let us recall that $\theta=\{A,W,B,C\}\in\Theta$, with $\Theta=\mathbb{R}^{d_\theta}$ where $d_\theta=N\times d+N\times N+N+d$ is be the dimension of the parameter space $\Theta$. If $\diamond$ and $\heartsuit$ are the labels of the parameters (i.e., $A^{ij},W^{ij},B^i$ or $C^i$), then

$$\begin{aligned}R_t&=(\theta_{t+1}-\theta_t)^\top H_\theta(\theta_t^*)(\theta_{t+1}-\theta_t)\\&=\sum_{\heartsuit,\diamond}(\theta_{t+1}^\heartsuit-\theta_t^\heartsuit)\frac{\partial^2L(\theta_t^*)}{\partial\heartsuit\partial\diamond}(\theta_j^\diamond-\theta_j^\diamond)\end{aligned}\tag{5.29}$$

By Corollary 3.17 we have that the second order partial derivatives $\frac{\partial^2 L(\theta_t^*)}{\partial \heartsuit \partial \diamondsuit}$ are uniformly bounded. Namely, there is some finite constant $C < \infty$ so that for all $i, j \in \{1, \cdots, d_\Theta\}$

$$\left\| \frac{\partial^2 L(\theta_t^*)}{\partial \heartsuit \partial \diamondsuit} \right\| \leq C. \tag{5.30}$$

This, together with the bound (3.62) and the assumption $\sum_{t=0}^{\infty} \alpha_t^2 < \infty$ yields

$$\sum_{t=j}^{T-1} |R_t| \leq C d_\theta \sum_{t=j}^{T-1} |\theta_{t+1} - \theta_t|_2^{\max} \leq C \sum_{t=j}^{T-1} \alpha_t^2 < \infty,$$

for some constant $C < \infty$ that may differ from line to line but is uniform in time. This completes the proof of the lemma. □

We are now ready to prove Proposition 5.1.

*Proof.* (of Proposition 5.1) The proof follows immediately by triangle inequality and using the bounds from Lemmas 5.2, 5.3, 5.4 and 5.5. □

# 6 Convergence as $t \to \infty$

We will now prove convergence of the RTRL algorithm as $t \to \infty$. Recall that

$$L(\theta_{t+1}) = L(\theta_t) - \alpha_t \nabla_\theta L(\theta_t)^\top \nabla_\theta L(\theta_t) + \alpha_t \nabla_\theta L(\theta_t)^\top (G_t - \nabla_\theta L(\theta_t)) + R_t. \tag{6.1}$$

Recalling Proposition 5.1 this representation suggests potential convergence of $\theta_t$ to critical points of $L(\theta)$. In particular, in this section we establish Theorem 6.1.

**Theorem 6.1.** *Assume that Assumptions 2.3, 2.10 and 2.14 hold. Then, we have that*

$$\lim_{t \to \infty} \mathbb{E} \|\nabla_\theta L(\theta_t)\|_2 = 0.$$

This theorem builds on some intermediate estimates that collectively lead to the statements for the liminf and the limsup of Lemmas 6.3 and 6.4 respectively that then directly yield the convergence theorem.

**Lemma 6.2.** *Assume that Assumptions 2.3, 2.10 and 2.14 hold.* $\mathbb{E}[L(\theta_t)]$ *converges to a finite value as* $t \to \infty$ *and* $\sum_{t=0}^{\infty} \alpha_t \mathbb{E}[\nabla_\theta L(\theta_t)^\top \nabla_\theta L(\theta_t)] < \infty$.

*Proof.* We adapt the proof from Lemma 1 in [15] to prove this lemma.

Let $i < T$. The evolution of the loss function from time step $i$ to $T$ satisfies:

$$\begin{aligned} L(\theta_T) &= L(\theta_i) - \sum_{t=i}^{T-1} \alpha_t \nabla_\theta L(\theta_t)^\top \nabla_\theta L(\theta_t) + \sum_{t=i}^{T-1} \alpha_t \left( u(\hat{H}_t; \theta_t) - \int_{\mathcal{R}} u(\mathsf{h}; \theta_t) P_{\theta_t}(\hat{H}_t, d\mathsf{h}) \right) + \sum_{t=i}^{T-1} R_t \\ &= L(\theta_i) - \sum_{t=i}^{T-1} \alpha_t \nabla_\theta L(\theta_t)^\top \nabla_\theta L(\theta_t) + Q_{i+1,T}^{1,A} + Q_{i+1,T}^{1,B} + Q_{i+1,T}^{2} + R_i^1 + \sum_{t=i}^{T-1} R_t. \end{aligned} \tag{6.2}$$

Taking an expectation of (6.2) and using the Cauchy-Schwarz inequality yields:

$$\begin{aligned}\mathbb{E}\Big[L(\theta_T)\Big] &= \mathbb{E}\Big[L(\theta_i)\Big] - \sum_{t=i}^{T-1}\alpha_t\mathbb{E}\Big[\nabla_\theta L(\theta_t)^\top \nabla_\theta L(\theta_t)\Big] \\ &+ \mathbb{E}\Big[|Q^{1,A}_{i+1,T}|\Big] + \mathbb{E}\Big[|Q^{1,B}_{i+1,T}|^2\Big]^{\frac{1}{2}} + \mathbb{E}\Big[|Q^2_{i+1,T}|\Big] + \mathbb{E}\Big[|R^1_i|\Big] \\ &+ \sum_{t=i}^{T-1}\mathbb{E}\Big[|R_t|\Big]. \end{aligned} \tag{6.3}$$

Then, applying the bounds from equation (5.8) and Lemmas 5.2, 5.5, 5.3 and 5.24, the evolution of the loss function satisfies the following bound:

$$\mathbb{E}\Big[L(\theta_T)\Big] \leq \mathbb{E}\Big[L(\theta_i)\Big] - \sum_{t=i}^{T-1}\alpha_t\mathbb{E}\Big[\nabla_\theta L(\theta_t)^\top \nabla_\theta L(\theta_t)\Big] + C_1\alpha_i + C_2\sum_{t=i}^{T-1}\alpha_t^2 + C_3\Big(\sum_{t=i}^{T-1}\alpha_t^2\Big)^{\frac{1}{2}}. \tag{6.4}$$

Since the second term on the right hand side of (6.4) is non-positive,

$$\mathbb{E}\Big[L(\theta_T)\Big] \leq \mathbb{E}\Big[L(\theta_i)\Big] + C_1\alpha_i + C_2\sum_{t=i}^{T-1}\alpha_t^2 + C_3\Big(\sum_{t=i}^{T-1}\alpha_t^2\Big)^{\frac{1}{2}}.$$

The above inequality implies that

$$\limsup_{T\to\infty}\mathbb{E}\Big[L(\theta_T)\Big] \leq \mathbb{E}\Big[L(\theta_i)\Big] + C_1\alpha_i + C_2\sum_{t=i}^{\infty}\alpha_t^2 + C_3\Big(\sum_{t=i}^{\infty}\alpha_t^2\Big)^{\frac{1}{2}} < \infty. \tag{6.5}$$

Since $\sum_{t=0}^{s}\alpha_t^2$ is monotone increasing in $s$ and $\sum_{t=0}^{\infty}\alpha_t^2 < C < \infty$, the monotone convergence theorem implies that it must converge to a finite value $L$ as $s \to \infty$. Then, $\sum_{t=i}^{\infty}\alpha_t^2 = L - \sum_{t=0}^{i-1}\alpha_t^2$ and $\lim_{i\to\infty}\sum_{t=i}^{\infty}\alpha_t^2 = 0$. Therefore,

$$\limsup_{T\to\infty}\mathbb{E}\Big[L(\theta_T)\Big] \leq \liminf_{i\to\infty}\mathbb{E}\Big[L(\theta_i)\Big] < \infty. \tag{6.6}$$

Therefore, since $L(\theta)$ is non-negative, $\mathbb{E}\Big[L(\theta_T)\Big]$ converges to a finite value. Furthermore, using the bound (6.4) and letting $T \to \infty$,

$$\begin{aligned}\sum_{t=0}^{\infty}\alpha_t\mathbb{E}\Big[\nabla_\theta L(\theta_t)^\top \nabla_\theta L(\theta_t)\Big] &\leq -\lim_{T\to\infty}\mathbb{E}\Big[L(\theta_T)\Big] + \mathbb{E}\Big[L(\theta_0)\Big] + C_1\alpha_0 + C_2\sum_{t=0}^{\infty}\alpha_t^2 + C_3\Big(\sum_{t=0}^{\infty}\alpha_t^2\Big)^{\frac{1}{2}} \\ &\leq \mathbb{E}\Big[L(\theta_0)\Big] + C_1\alpha_0 + C_2\sum_{t=0}^{\infty}\alpha_t^2 + C_3\Big(\sum_{t=0}^{\infty}\alpha_t^2\Big)^{\frac{1}{2}} \\ &\leq C < \infty. \end{aligned}$$

□

**Lemma 6.3.** *Assume that Assumptions 2.3, 2.10 and 2.14 hold. We have that*

$$\liminf_{t\to\infty}\mathbb{E}\|\nabla_\theta L(\theta_t)\|_2 = 0. \tag{6.7}$$

*Proof.* The proof of this lemma is standard and follows directly from Lemma 6.2. Indeed, let us assume that there is some constant $\eta > 0$ so that for some $\tau > 0$ and for all $t \geq \tau$ we have that $\mathbb{E}\|\nabla_\theta L(\theta_t)\|_2^2 \geq \eta$. Then, we naturally have that

$$\sum_{t=\tau}^{\infty} \alpha_t \mathbb{E}\|\nabla_\theta L(\theta_t)\|_2^2\| \geq \eta \sum_{t=\tau}^{\infty} \alpha_t = \infty, \tag{6.8}$$

which contradicts the statement of Lemma 6.2. This concludes the proof of the lemma. □

**Lemma 6.4.** *Assume that Assumptions 2.3, 2.10 and 2.14 hold. We have that*

$$\limsup_{t\to\infty} \mathbb{E}\|\nabla_\theta L(\theta_t)\|_2 = 0. \tag{6.9}$$

*Proof.* The proof of this result is classical and largely follows the proof of Proposition 1 in [15] modulo some necessary adjustments due to the differences on the setup. We present the full details for completeness.

Let us assume that $\limsup_{t\to\infty} \mathbb{E}\|\nabla_\theta L(\theta_t)\|_2 > 0$. This means that there is some $\eta > 0$ such that $\mathbb{E}\|\nabla_\theta L(\theta_t)\|_2 < \eta/2$ for infinitely many $t$ and $\mathbb{E}\|\nabla_\theta L(\theta_t)\|_2 > \eta$ for infinitely many $t$. This means that there is an infinite subset of integers $B \subset \mathbb{N}_+$ so that for all $t \in B$ there is some $\hat{t} > t$ with the properties

$$\begin{aligned} \mathbb{E}\|\nabla_\theta L(\theta_t)\|_2 &< \frac{\eta}{2} \\ \mathbb{E}\|\nabla_\theta L(\theta_{\hat{t}})\|_2 &> \eta \\ \frac{\eta}{2} < \mathbb{E}\|\nabla_\theta L(\theta_s)\|_2 &< \eta, \text{ for } s \in (t, \hat{t}). \end{aligned}$$

Clearly $\hat{t}$ depends on $t$, but we do not show this explicitly in the notation. By Corollary 3.17, $\nabla_\theta L(\theta)$ is globally Lipschitz in $\ell_2$ norm, so we have

$$\begin{aligned} \mathbb{E}\|\nabla_\theta L(\theta_{t+1})\|_2 - \mathbb{E}\|\nabla_\theta L(\theta_t)\|_2 &\leq \mathbb{E}\|\nabla_\theta L(\theta_{t+1}) - \nabla_\theta L(\theta_t)\|_2 \\ &\leq L_{\nabla L}\|\theta_{t+1} - \theta_t\|_2 \\ &\leq L_{\nabla L} C \alpha_t, \end{aligned}$$

where $L_{\nabla L}$ denotes the global Lipschitz constant of $\nabla L(\theta)$ and where we used (3.62) in the last line.

Let now $t \in B$ large enough so that $L_{\nabla L} C \alpha_t < \frac{\eta}{4}$. Then, we shall have that $\mathbb{E}\|\nabla_\theta L(\theta_t)\|_2 > \frac{\eta}{4}$, otherwise the condition $\mathbb{E}\|\nabla_\theta L(\theta_{t+1})\|_2 > \frac{\eta}{2}$ will be violated. Hence, for all $t \in B$ we shall have that $\mathbb{E}\|\nabla_\theta L(\theta_t)\|_2 > \frac{\eta}{4}$. Next, with that bound in mind, we have the following estimate for all $t \in B$

$$\begin{aligned} \frac{\eta}{2} &\leq \mathbb{E}\|\nabla_\theta L(\theta_{\hat{t}})\|_2 - \mathbb{E}\|\nabla_\theta L(\theta_t)\|_2 \\ &\leq \mathbb{E}\|\nabla_\theta L(\theta_{\hat{t}}) - \nabla_\theta L(\theta_t)\|_2 \\ &\leq L_{\nabla L}\|\theta_{\hat{t}} - \theta_t\|_2 \\ &\leq L_{\nabla L} C \sum_{s=t}^{\hat{t}-1} \alpha_s, \end{aligned}$$

which then implies that

$$\liminf_{t\to\infty} \sum_{s=t}^{\hat{t}-1} \alpha_s \geq \frac{\eta}{2 L_{\nabla L} C} > 0. \tag{6.10}$$

Now, going back to (6.1) we get

$$\mathbb{E}\left[L(\theta_{\hat{t}})\right] = \mathbb{E}\left[L(\theta_t)\right] - \sum_{s=t}^{\hat{t}-1} \alpha_s \mathbb{E}\left[\|\nabla_\theta L(\theta_s)\|_2^2\right] + \sum_{s=t}^{\hat{t}-1} \mathbb{E}\left[\alpha_s \nabla_\theta L(\theta_s)^\top (G_s - \nabla_\theta L(\theta_s)) + R_s\right]$$
$$\leq \mathbb{E}\left[L(\theta_t)\right] - \left(\frac{\eta}{4}\right)^2 \sum_{s=t}^{\hat{t}-1} \alpha_s + \sum_{s=t}^{\hat{t}-1} \mathbb{E}\left[\alpha_s \nabla_\theta L(\theta_s)^\top (G_s - \nabla_\theta L(\theta_s)) + R_s\right].$$

This then gives the inequality

$$\sum_{s=t}^{\hat{t}-1} \alpha_s \leq \frac{16}{\eta^2}\left[\mathbb{E}\left[L(\theta_t)\right] - \mathbb{E}\left[L(\theta_{\hat{t}})\right] + \sum_{s=t}^{\hat{t}-1} \mathbb{E}\left[\alpha_s \nabla_\theta L(\theta_s)^\top (G_s - \nabla_\theta L(\theta_s)) + R_s\right]\right].$$

Using now the fact that $\mathbb{E}[L(\theta_t)]$ converges to a finite value as $t \to \infty$ via Lemma 6.2 and the uniform boundedness of $\sum_{s=0}^{\infty} \mathbb{E}\left[\alpha_s \nabla_\theta L(\theta_s)^\top (G_s - \nabla_\theta L(\theta_s)) + R_s\right]$, which follows from (5.4) and Proposition 5.1, we obtain that

$$\lim_{t\to\infty, t\in B} \sum_{s=t}^{\hat{t}-1} \alpha_s = 0,$$

which contradicts (6.10). Hence, we shall indeed have that $\limsup_{t\to\infty} \mathbb{E}\|\nabla_\theta L(\theta_t)\|_2 = 0$ as desired. □

# 7 Numerical Examples

In this section, we compare the numerical performance of the RTRL algorithm with the more widely-used TBPTT algorithm for several examples. Section 7.1 evaluates the performance of RTRL for linear RNNs on synthetic data. Section 7.2 evaluates RTRL performance for single-layer RNNs (with the Elman architecture) on synthetic data. A comparison of RTRL and TBPTT for neural ODEs is provided in Section 7.3. RTRL is compared with TBPTT for several different small-scale RNN architectures for natural language processing (NLP) on a sequence of length 1 million characters in Section 7.4. Section 7.5 compares RTRL with TBPTT on financial time series data for the order book of a stock.

In each of these cases, we focus on relatively small models (in terms of the number of parameters and hidden units) when comparing RTRL and TBPTT. Of course, deep learning has found that generally large models can perform better (e.g., millions of parameters), where RTRL would not be computationally tractable. Our objective here is therefore limited to evaluating whether, for a series of fixed problems, RTRL performs better than TBPTT. Improved computational performance would motivate (A) further research into computationally tractable approximations of RTRL for large-scale problems and (B) the application of RTRL to problems where small or medium-scal models are appropriate (i.e., limited or noisy data where large models may overfit).

## 7.1 Linear RNNs

Consider the linear RNN

$$\begin{aligned} S_{t+1} &= S_t + WS_t\Delta + BX_t\sqrt{\Delta}, \\ \hat{Y}_t &= AS_{t+1}, \\ L_T(\theta) &= \frac{1}{T}\sum_{t=1}^{T}(\hat{Y}_t - Y_t)^2, \end{aligned} \tag{7.1}$$

where $Y_t$ is generated from the process

$$\begin{aligned} S^*_{t+1} &= S^*_t + W^* S^*_t \Delta + B^* X_t \sqrt{\Delta}, \\ Y_t &= A^* S^*_{t+1}. \end{aligned} \tag{7.2}$$

(7.1) can be considered a (linear) neural SDE with time step size $\Delta$ or, alternatively, it can be re-written in the standard form of a linear RNN by recognizing that $S_{t+1} = \bar{W} S_t + \bar{B} X_t$ where $\bar{W} = I - W\Delta$ and $\bar{B} = B\sqrt{\Delta}$. To ensure ergodicity of $S_t$, we would need to constrain the $d \times d$ matrix $W$ to be negative definite (assuming a sufficiently small time step size $\Delta$). There are various approaches which could be implemented. $W = -\theta^\top \theta$ where $\theta$ is a $d \times d$ matrix parameters would guarantee that $W$ is negative semi-definite. A slightly more mathematically elegant construction of a negative definite matrix is $W = -\exp(\theta^\top \theta)$. In the following examples, we do not however impose any constraints on the matrix $W$ and simply let the RTRL/TBPTT algorithms try to directly learn an appropriate matrix parameter.

In the following numerical example, the data $X_t$ is a $2 \times 1$ vector with an i.i.d. Gaussian distribution, $W$ is a $10 \times 10$ matrix, $B$ is a $10 \times 2$ matrix, $A$ is a $1 \times 10$ vector, and $\Delta = 10^{-2}$. The objective is to train the parameters $\theta = (W, B, A)$ to minimize the error $L_T(\theta)$. The true parameters $\theta^* = (W^*, B^*, A^*)$, which are used to generate the data $Y_t$, are randomly initialized. $W^*$ is constructed from a Wishart distribution such that it is positive definite while $B^*$ and $A^*$ have i.i.d. standard normal elements. Before each training run, the model parameters $\theta$ are randomly initialized. The time step size $\Delta = 10^{-2}$.

The specific construction for the Wishart distribution of $W^*$ is:

$$W^* = \sum_{i=1}^{n} G^{(i)} (G^{(i)})^\top, \tag{7.3}$$

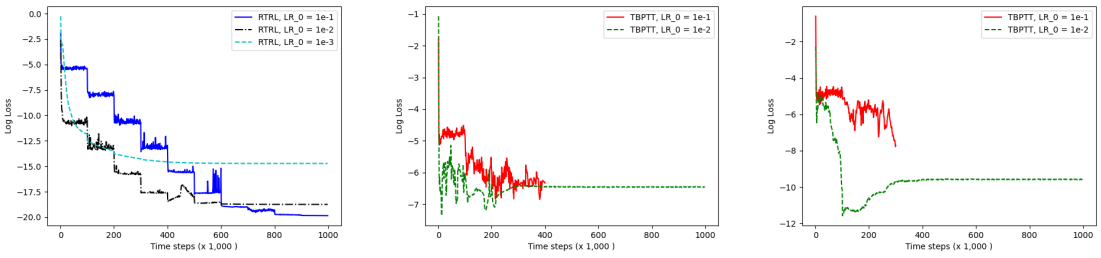
where $G^{(i)}$ is a $10 \times 1$ i.i.d. standard Gaussian vector and $n = 20$.

We compare the standard TBPTT and RTRL algorithms for training the parameters $\theta$ using a mini-batch size of $10^3$. Optimization updates are performed using the RMSprop algorithm with an initial learning rate magnitude $\mathrm{LR}_0$, which is gradually reduced according to a learning rate schedule. Figure 1 compares RTRL and TBPTT for several different initial learning rate magnitudes and choices of truncation length $\tau$ for TBPTT. Typically, RTRL performs better than TBPTT. In several cases, TBPTT becomes unstable (hence the log-loss $\log L_T(\theta)$ is only plotted for a few training iterations in some of the figures). In particular, TBPTT seems to require a very carefully selected choice of learning rate magnitude and truncation $\tau$ to achieve good performance. Increasing the standard deviation of the data sequence $X_t$ by a factor of 4, we re-train the linear RNN with the RTRL and TBPTT algorithms in Figure 2. RTRL outperforms TBPTT, typically achieving a much smaller loss.

## 7.2 Single-layer RNNs

We now consider a single-layer RNN with the standard Elman network architecture. Specifically,

$$\begin{aligned} S_{t+1} &= \sigma(W S_t + B X_t), \\ \hat{Y}_t &= A S_{t+1}, \\ L_T(\theta) &= \frac{1}{T} \sum_{t=1}^{T} (\hat{Y}_t - Y_t)^2, \end{aligned} \tag{7.4}$$



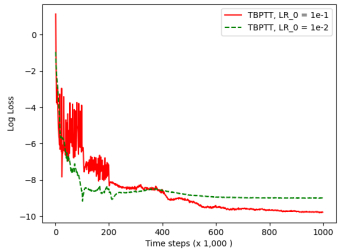

(a) RTRL



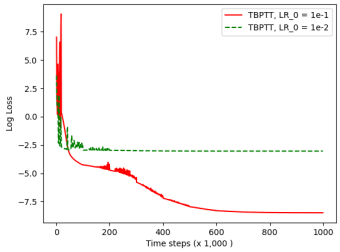

(b) TBPTT ($\tau = 1$)



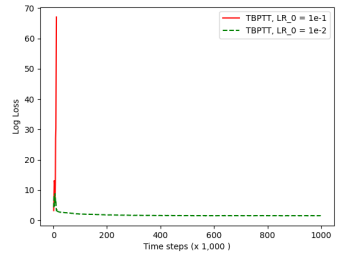

(c) TBPTT ($\tau = 2$)




(d) TBPTT ($\tau = 10$)

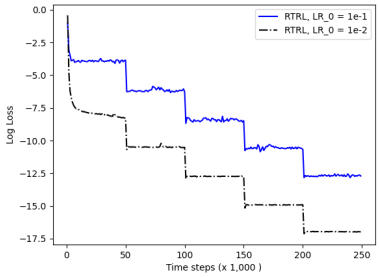



(e) TBPTT ($\tau = 100$)

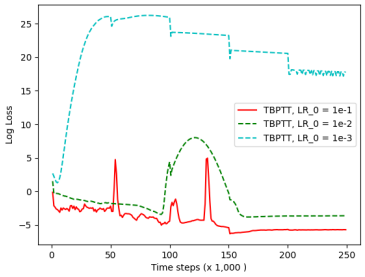



(f) TBPTT ($\tau = 1000$)

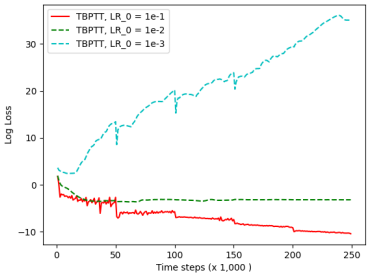
Figure 1: Clockwise: RTRL, TBPTT ($\tau = 1$), TBPTT ($\tau = 2$), TBPTT ($\tau = 10$), TBPTT ($\tau = 100$), and TBPTT ($\tau = 1,000$).



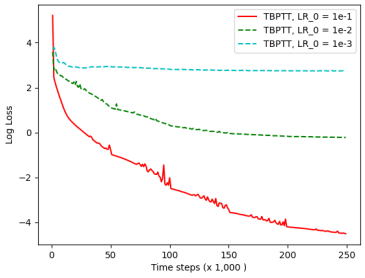

(a) RTRL




(b) TBPTT ($\tau = 2$)




(c) TBPTT ($\tau = 10$)




(d) TBPTT ($\tau = 100$)

Figure 2: Clockwise: RTRL, TBPTT ($\tau = 2$), TBPTT ($\tau = 10$), and TBPTT ($\tau = 100$).

where $Y_t$ is generated from the process

$$\begin{aligned} S^*_{t+1} &= \sigma\big(\bar{W}^* S^*_t + B^* X_t\big), \\ Y_t &= A^* S^*_{t+1}, \end{aligned} \tag{7.5}$$

where the activation function $\sigma(\cdot)$ is the tanh function and the weight matrix $\bar{W}^*$ is constructed as $(I - cW^*)$ with $c = 10^{-3}$. The true parameters $W^*, B^*$, and $A^*$ are randomly generated in the same way as in the previous section except $B^*_{i,j}$ has standard deviation $10^{-1}$. $X_t$ is a $2 \times 1$ vector with an i.i.d. Gaussian distribution. For the case of ReLU activation functions (i.e., $\sigma(z) = \max(z, 0)$), Figure 4 displays convergence results for the TBPTT algorithm while Figure 3 presents the training results for the RTRL algorithm. RTRL outperforms TBPTT for ReLU activation functions.

We next consider a numerical comparison where the activation function $\sigma(z) = \tanh(z)$. The matrix $\bar{W}^*$ is again constructed as $(I - cW^*)$ with $c = 10^{-3}$. The TBPTT results are presented in Figure 6 while the RTRL results are displayed in Figure 5. RTRL outperforms TBPTT, with faster convergence and a lower overall value for the objective function. Finally, we consider $\sigma(z) = \tanh(z)$ where the matrix $\bar{W}^* = (I - cW^*)$ is constructed with a different value $c = 10^{-2}$. The corresponding RTRL and TBPTT results are displayed in Figures 7 and 8, respectively. RTRL outperforms TBPTT in this case.

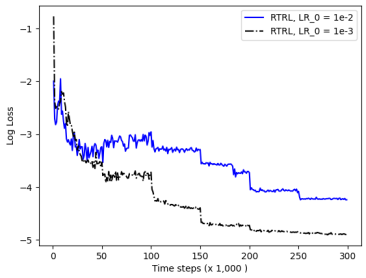
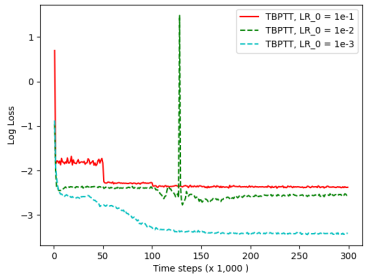



Figure 3: RTRL algorithm for ReLU activation functions.

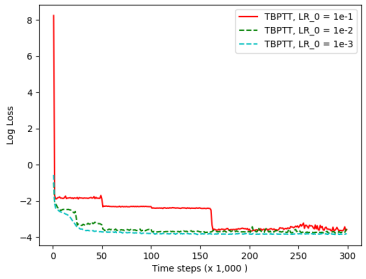
### 7.3 Neural ODEs/SDEs

We now compare the RTRL and TBPTT algorithms for training a (discretized) neural SDE:

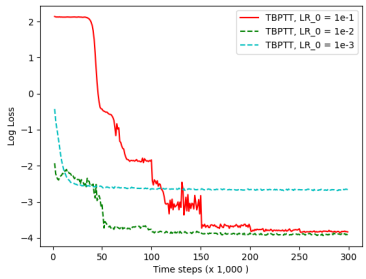
$$\begin{aligned} S_{t+1} &= S_t - \sigma(US_t + BX_t)\Delta - CS_t\Delta, \\ \hat{Y}_t &= AS_{t+1}, \\ L_T(\theta) &= \frac{1}{T}\sum_{t=1}^{T}(\hat{Y}_t - Y_t)^2, \end{aligned} \tag{7.6}$$



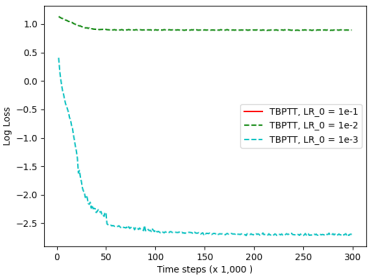

(a) TBPTT ($\tau = 1$)




(b) TBPTT ($\tau = 2$)




(c) TBPTT ($\tau = 10$)




(d) TBPTT ($\tau = 1000$)

Figure 4: Activation function is a ReLU unit.

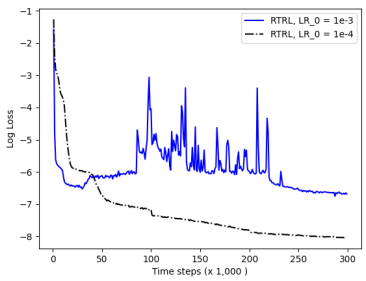
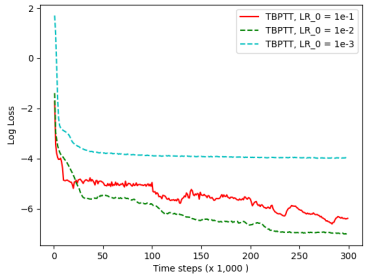
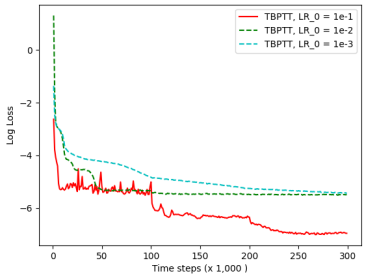


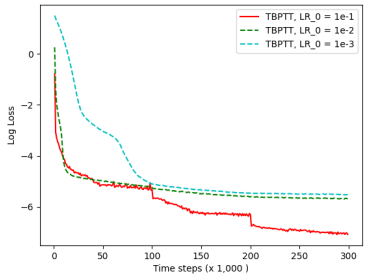

Figure 5: RTRL algorithm for tanh activation functions with $c = 10^{-3}$.



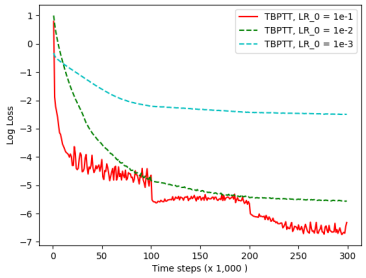

(a) TBPTT ($\tau = 1$)




(b) TBPTT ($\tau = 2$)




(c) TBPTT ($\tau = 10$)




(d) TBPTT ($\tau = 1000$)

Figure 6: Activation function is a tanh unit with $c = 10^{-3}$.

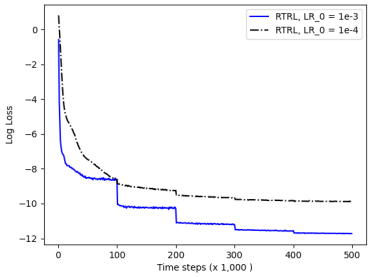
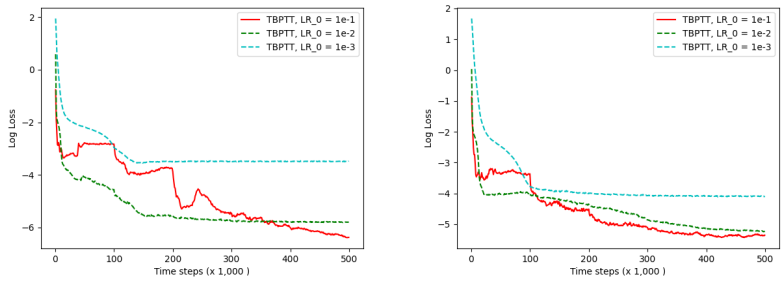
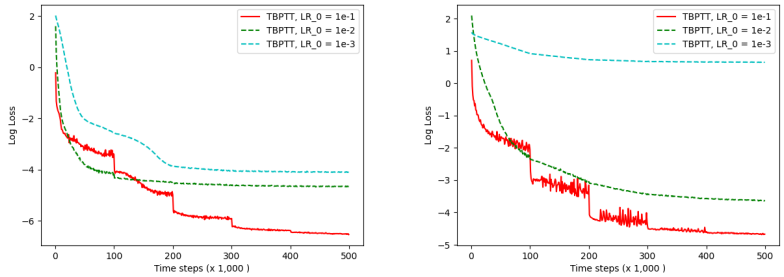



Figure 7: RTRL algorithm for tanh activation functions with $c = 10^{-2}$.




(a) TBPTT ($\tau = 1$) (b) TBPTT ($\tau = 2$)




(c) TBPTT ($\tau = 10$) (d) TBPTT ($\tau = 1000$)

Figure 8: Activation function is a tanh unit with $c = 10^{-2}$.

where $C = \exp(W^\top W)$ is positive definite, the parameters to be optimized are $\theta = (U, B, W)$, and $Y_t$ is generated from the process

$$\begin{aligned} S^*_{t+1} &= S^*_t - \sigma(\bar{U}^* S^*_t + B^* X_t)\Delta - W^* S_t \Delta, \\ Y_t &= A^* S^*_{t+1}, \end{aligned} \tag{7.7}$$

where $W^*$ is a positive definite matrix generated from the Wishart distribution. Figure 9 compares the performance of the TBPTT and RTRL algorithms for training the parameters $\theta$.

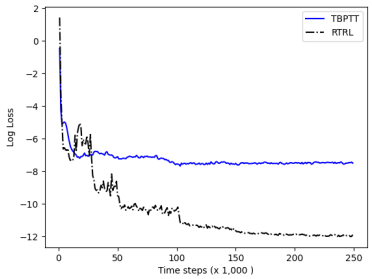



Figure 9: Comparison of TBPTT and RTRL for training a neural ODE.

### 7.4 Natural Language Processing

We now compare RTRL and TBPTT for predicting the next character in a sequence of 1 million characters from Shakespeare [22]. The objective function is the cross-entropy error. Figure 10 presents the comparisons for several different architectures, including: a GRU-type architecture with 10 and 20 units, an Elman network with tanh activation functions 10 and 50 units, and an Elman network with ELU activation functions with 10 and 50 units. RTRL does not consistently perform better than TBPTT, although it does perform better in several notable cases, including the GRU network with 20 units. Since the results are for a specific dataset with relatively small model sizes (due to the computational cost of RTRL), it is difficult to draw general conclusions except that it may be worthwhile, for a specific dataset, to implement and compare both RTRL and TBPTT to determine which yields the most accurate model.

### 7.5 Financial Time Series Data

Finally, we compare RTRL and TBPTT for training a single-layer Elman recurrent network on financial time series data. The dataset is order book data from the stock Amazon consisting of a time series of approximately 180 million order book events from January 2, 2014 - September 1, 2017. The objective function is the cross-entropy error to classify whether the mid-price increases, decreases, or stays the same at the next event in the time series. Recurrent neural network classification models for price moves have been previously trained on order book data using the TBPTT algorithm [23]. In the example here, the recurrent network has 25 hidden units with tanh() activation functions. The input variables to the recurrent neural network are the previous price change and the current order book queue imbalance [24]. A softmax layer is applied to its output to produce a probability distribution over the three classes (price increases, decreases, or stays the same). Figure 11 compares the training results from TBPTT and RTRL.

## Acknowledgements

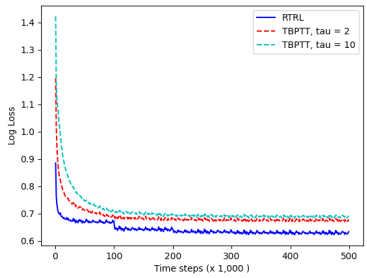
This article is part of the project "DMS-EPSRC: Asymptotic Analysis of Online Training Algorithms in Machine Learning: Recurrent, Graphical, and Deep Neural Networks" (NSF DMS-2311500). Samuel Lam's research is also supported by the EPSRC Centre for Doctoral Training in Mathematics of Random Systems: Analysis, Modelling and Simulation (EP/S023925/1).

## A Bounds for the Derivatives of Sigmoid Function

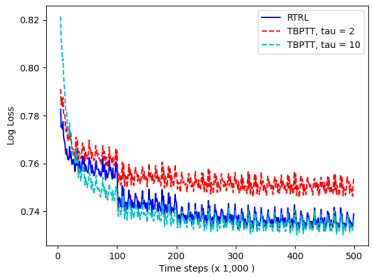
*Proof of Lemma 2.9.* Since $e^{-y} > 0$, we have $0 < \sigma(y) < 1$, which leads to the first bound. The second bound follows from:

$$\sigma'(y) = p_1(\sigma(y)), \quad p_1(x) = x(1-x) = x - x^2 = -\left(x - \frac{1}{2}\right)^2 + \frac{1}{4}.$$

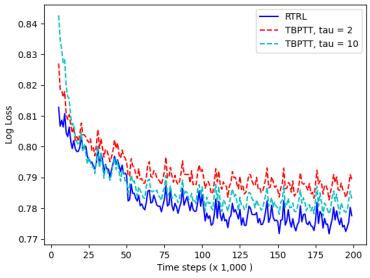
Since $0 < \sigma(y) < 1$, it follows that $0 < \sigma'(y) < 1/4$, with maximum of $\sigma'(y)$ attained at $\sigma(y) = 1/2 \iff y = 0$. This leads to the second bound.



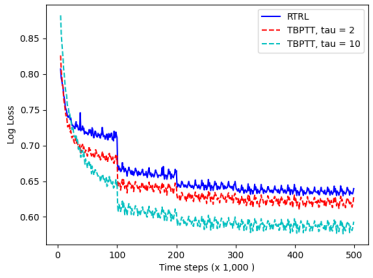

(a) GRU (10 hidden units)



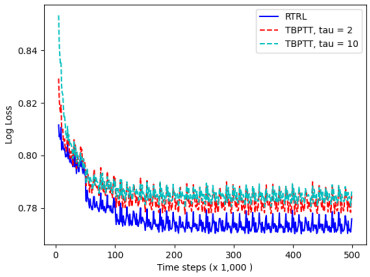

(b) GRU (20 hidden units)



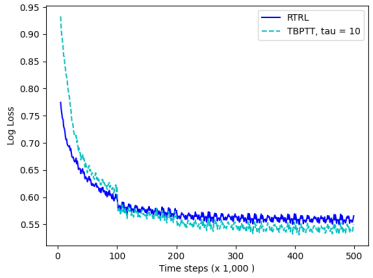

(c) Elman RNN with 10 tanh hidden units




(d) Elman RNN with 50 tanh hidden units




(e) Elman RNN with 10 ELU hidden units




(f) Elman RNN with 50 ELU hidden units

Figure 10: First row: a GRU network (10 hidden units) and a GRU network (20 hidden units). Second row: Elman RNN with 10 tanh hidden units and 50 tanh hidden units. Third row: Elman RNN with 10 ELU hidden units and 50 ELU hidden units.

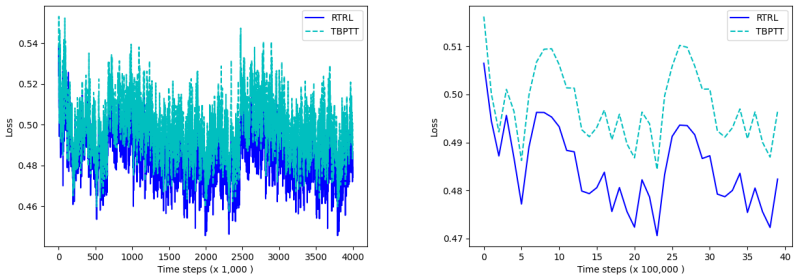



Figure 11: Left: Loss averaged over $1,000$ time steps for each "training epoch". Right: Loss averaged over $100,000$ time steps for each "training epoch" to provide a less noisy comparison.

Differentiate again yields

$$\begin{aligned}\sigma''(y) &= \sigma'(y) - 2\sigma(y)\sigma'(y)\\ &= \sigma(y) - \sigma(y)^2 - 2\sigma(y)^2 + 2\sigma(y)^3\\ &= p_2(\sigma(y)),\end{aligned}$$

where

$$p_2(x) = 2x^3 - 3x^2 + x = 2\left(x - \frac{1}{2}\right)^3 - \left(x - \frac{1}{2}\right).$$

The derivative of $p_2(x)$ is

$$p_2'(x) = 6x^2 - 6x + 1 = 6\left(x - \frac{1}{2}\right)^2 - \frac{1}{2},$$

which has zeroes at

$$x_*^{\pm} = \frac{1}{2} \pm \frac{\sqrt{3}}{6}, \quad x_*^+ \approx 0.789, \quad x_*^- \approx 0.211.$$

Checking $p_2(0) = p_2(1) = 0$, one could confirm that $\sigma''(y)$ attains maxmimum at $\sigma(y) = p^+$ and minimum at $\sigma(y) = p^-$. One could then show that $p_2(x_*^+) \le \sigma''(y) \le p_2(x_*^-)$, where

$$p_2(x_*^-) = -p_2(x_*^+) = \frac{\sqrt{3}}{18} < \frac{1.8}{18} = \frac{1}{10}.$$

Differentiate again yields

$$\begin{aligned}\sigma'''(y) &= 6\sigma(y)^2\sigma(y)(1-\sigma(y)) - 6\sigma(y)^2(1-\sigma(y)) + \sigma(y)(1-\sigma(y))\\ &= 6\sigma(y)^3 - 6\sigma(y)^4 - 6\sigma(y)^2 + 6\sigma(y)^3 + \sigma(y) - \sigma(y)^2\\ &= p_3(\sigma(y)),\end{aligned}$$

where

$$p_3(x) = -6x^4 + 12x^3 - 7x^2 + x = -6\left(x - \frac{1}{2}\right)^4 + 2\left(x - \frac{1}{2}\right)^2 - \frac{1}{8}.$$

The polynomial $p_3(x)$ has derivative

$$p_3'(x) = -24x^3 + 36x^2 - 14x + 1 = -24\left(x - \frac{1}{2}\right)^3 + 4\left(x - \frac{1}{2}\right),$$

which has zeros at $x_* = 1/2$ **and**

$$x_\pm = \frac{1}{2} \pm \sqrt{\frac{1}{6}}, \quad x_+ \approx 0.908, \quad x_- \approx 0.092.$$

Checking again $p_3(0) = p_3(1) = 0$, and noting that

$$p_3(x_+) = p_3(x_-) = \frac{1}{24}, \quad m(x_*) = -\frac{1}{8},$$

one could confirm that $\sigma'''(y)$ attains maximum at $\sigma(y) = x_\pm$ and minimum at $\sigma(y) = x_* = 1/2 \equiv y = 0$. Therefore $-1/8 = p_3(x_*) \le \sigma'''(y) \le p_3(x_\pm) = 1/24$, which leads to the final bound. □

# B Important Inequalities

We indicate the application of the following inequalities

- the Cauchy-Schwarz inequality: for any $a_k, b_k \in \mathbb{R}$

$$\left(\sum_{k=1}^n a_k b_k\right)^2 \le \left(\sum_{k=1}^n a_k^2\right)\left(\sum_{k=1}^n b_k^2\right), \tag{B.1}$$

  as well as its special case

$$\left(\sum_{k=1}^n a_k\right)^2 \le n\left(\sum_{k=1}^n a_n^2\right), \tag{B.2}$$

- the Young's inequality: for any $\epsilon > 0$,

$$|ab| \le \frac{\epsilon}{2}a^2 + \frac{1}{2\epsilon}b^2, \tag{B.3}$$

- the Jensen inequality: for any convex functions $\varphi$, and sequence $(w_k)$ such that $\sum_k w_k = 1$, we have

$$\varphi\left(\sum_{k=1}^n w_k x_k\right) \le \sum_{k=1}^n w_k \varphi(x_k) \tag{B.4}$$

Many proofs of the technical lemmas also involves the study of a sequence $(a_k)_{k\ge 0}$. If the sequence satisfies the following recursive inequality:

$$a_k \le M_1 a_{k-1} + M_2,$$

for some $M_1, M_2 \ge 0$, then

- for $M_1 < 1$:

$$a_k \le M_1^k a_0 + \frac{1 - M_1^k}{1 - M_1} M_2 \le M_1^k a_0 + \frac{1}{1 - M_1} M_2, \tag{B.5}$$

  which imply

$$a_0 \le \frac{M_2}{1 - M_1} \implies a_k \le \frac{M_2}{1 - M_1},$$

- and for $M_1 > 1$,

$$a_k \le M_1^k a_0 + \frac{M_1^k - 1}{M_1 - 1} M_2 \le M_1^k \left(a_0 + \frac{M_2}{M_1 - 1}\right). \tag{B.6}$$

# References


[1] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 1986;323:533-6. Available from: https://api.semanticscholar.org/CorpusID:205001834.

[2] Werbos PJ. Backpropagation through time: what it does and how to do it. Proceedings of the IEEE. 1990;78(10):1550-60.

[3] Williams RJ, Zipser D. A Learning Algorithm for Continually Running Fully Recurrent Neural Networks. Neural Computation. 1989;1(2):270-80.

[4] Williams RJ, Zipser D. Experimental Analysis of the Real-time Recurrent Learning Algorithm. Connection Science. 1989;1(1):87-111.

[5] Robinson T. Dynamic error propagation networks. PhD thesis, University of Cambridge, UK. 1989.

[6] Robinson T, Fallside F. The utility driven dynamic error propagation network. University of Cambridge, Department of Enginnering, UK. 1987.

[7] Menick J, Elsen E, Evci U, Osindero S, Simonyan K, Graves A. Practical Real Time Recurrent Learning with a Sparse Approximation. In: International Conference on Learning Representations; 2021. Available from: https://openreview.net/forum?id=q3KSThy2GwB.

[8] Irie K, Gopalakrishnan A, Schmidhuber J. Exploring the Promise and Limits of Real-time Recurrent Learning. In: ICLR; 2024. .

[9] Tallec C, Ollivier Y. Unbiased Online Recurrent Optimization. International Conference on Learning Representations (ICLR), Vancouver, Canada. 2018.

[10] Mujika A, Meier F, Steger A. Approximating Real-Time Recurrent Learning with Random Kronecker Factors. In: Neural Information Processing Systems; 2018. Available from: https://api.semanticscholar.org/CorpusID:44111521.

[11] Benzing F, Gauy M, Mujika A, Martinsson A, Steger A. Optimal Kronecker-sum approximation of real time recurrent learning. In: International Conference on Machine Learning; 2019. p. 604-13.

[12] Silver D, Goyal A, Danihelka I, Hessel M, van Hasselt H. Learning by Directional Gradient Descent. In: International Conference on Learning Representations; 2022. Available from: https://openreview.net/forum?id=5i7lJLuhTm.

[13] Schmidhuber J. A Fixed Size Storage O(n3) Time Complexity Learning Algorithm for Fully Recurrent Continually Running Networks. Neural Computation. 1992;4(2):243-8.

[14] Williams RJ, Zipser D. Gradient-based learning algorithms for recurrent networks and their computational complexity; 1995.

[15] Bertsekas D, Tsitsiklis J. Gradient convergence in gradient methods via errors. SIAM Journal of Optimization. 2000;10(3):627-42.

[16] Massé PY, Ollivier Y. Convergence of Online Adaptive and Recurrent Optimization Algorithms. arXiv: 2005/05645. 2020.

[17] Villani C. Optimal transport: old and new. Springer, Berlin; 2009.

[18] Bogachev VI, Kolesnikov AV. The Monge–Kantorovich problem: achievements, connections, and perspectives. Russian Mathematical Surveys. 2012;67:785-890.

[19] Dobrushin R, Groeneboom P, Ledoux M. Lectures on Probability Theory and Statistics: Ecole d'Eté de Probabilités de Saint-Flour XXIV—1994. vol. 1648 of Lecture Notes in Mathematics. Berlin, Heidelberg: Springer Berlin Heidelberg; 2006.

[20] Rudin W. Principles of Mathematical Analysis. 3rd ed. McGraw Hill; 1953.

[21] Lam SCH, Sirignano J, Spiliopoulos K. Kernel Limit of Recurrent Neural Networks Trained on Ergodic Data Sequences; 2023.

[22] Karpathy A. The Unreasonable Effectiveness of Recurrent Neural Networks. http://karpathygithubio/2015/05/21/rnn-effectiveness/. 2015.

[23] Sirignano J, Cont R. Universal Price Formation in Financial Markets: Insights from Deep Learning. Quantitative Finance. 2019;9.

[24] Gould MD, Bonart J. Queue Imbalance as a One-Tick-Ahead Price Predictor in a Limit Order Book. Market Microstructure and Liquidity. 2016;2.